Machine Learning based approaches for the Gene-Based Diagnosis of Parkinson's Disease

A Thesis

submitted for the award of the degree of

Doctor of Philosophy

in

Computer Science and Engineering Department

Submitted by

Priya Arora
(Reg no: 901403029)

Under the Guidance of

Dr. Ashutosh Mishra
Assistant Professor

Dr. Avleen Malhi
Senior Lecturer in Data Science and AI
Bournemouth University, UK

Thapar Institute of Engineering and Technology, Patiala,
Punjab - 147004, India
February 2023

Certificate

This is to certify that the thesis entitled, "**Machine Learning based approaches for the Gene-Based Diagnosis of Parkinson's Disease** ", in partial fulfillment of the requirements for the award of the degree of DOCTOR OF PHILOSOPHY submitted in the Computer Science and Engineering Department (CSED), Thapar Institute of Engineering and Technology (TIET), Patiala, Punjab, is an authentic record of my own work carried out under the supervision of Dr. Ashutosh Mishra and Dr. Avleen Malhi. I have cited the reference about the text(s)/figure(s)/table(s) from where they have been taken.

The matter presented in this thesis has not been submitted either in-part or full to any other University or Institute for the award of any other degree.

(Priya Arora)
Registration No. 901403029

This is to certify that the above statements made by the candidate are correct and true to the best of my knowledge.

Verified by:

(Dr. Ashutosh Mishra)
Assistant Professor
Computer Science and Engineering Department
Thapar Institute of Engineering and Technology, Patiala, Punjab, India

(Dr. Avleen Malhi)
Senior Lecturer in Data Science and AI
Department of Computing and Informatics
Bournemouth University, UK

Acknowledgements

I am highly grateful to God for his inspiration and blessings throughout my journey of the Ph.D. work. This Ph.D. has been a truly life-changing experience for me and it would not have been possible to do without the support and encouragement of numerous people including my family, supervisors, colleagues, friends and various institutions. I have had the great luck to work and collaborate with a lot of inspiring, competent and nice people who contributed in many ways to the success of this study. It is a pleasant task to use this space to thank them.

At this moment of accomplishment, first of all I am indebted to my supervisors, Dr. Ashutosh Mishra and Dr. Avleen Malhi, under whose guidance, I successfully overcome many difficulties and learnt a lot. Dr. Malhi steered me through this journey with her valuable advice, positive criticism, stimulating discussions and consistent encouragement. She took care to shine light of knowledge, when I was groping in the dark of ignorance. If I will stand proud of my achievements then undeniably both are the main creditors. It had been my privilege to work under their guidance.

It is my privilege to thank Prof. Prakash Gopalan, Director, Thapar Institute of Engineering and Technology, Patiala for providing me resources in this institution. I am also grateful to Dr. Shalini Batra, Head of Department of Computer Science and also my doctoral committee members for providing me the necessary administrative assistance,kind help and support as and when required in completion of the thesis work. A very special thanks to Dr. Prashant Singh Rana, one of my doctoral committee members, for his meticulous attention towards my proceedings and devoted time. His ideas has enabled me to make this journey a success. I am also grateful to Dr. Kulbir Singh for monitoring my work and suggestions to enhance the quality of work. I would also like to acknowledge Dr. O.P. Pandey, Sr. Professor, School of Physics and Materials Science for the indispensable academic and technical support. I express my gratitude for providing valuable suggestions for improvement of my thesis work time to time.

My in-laws family deserves endless gratitude. Thank you to my partner Mr.Varun Sadana for constantly listening to me, for cracking jokes when things became too serious, even after long days at work and during difficult times. Also, I would like to thank my father-in-law Sh.Dharampal Sadana and mother-in-law Mrs.Neelam Sadana for giving

me strength and safe keeping my children (Jenia and Tanish) when I wasn't able to care them mentally or physically.

My mother-in-law deserves special thanks during this phase of life, she has always been supportive and has encouraged me in my academic growth from the beginning of my Ph.D. journey. Sometimes, when things got too serious, she was there to heel and has also given her life experiences to boost up my power. Her company will always be special and will be always remembered.

Immense gratitude and thanks to my Parents, Prince, Ravina and Ritika for their support, love and affection for always with me till achieving my end goals.

At the end, I would also like to thank my fellow research scholars, especially Dr.Gagandeep Singh Aujla, Dr.Anish Jindal, Dr.Baljit Kaur, Dr.Loveleen Kaur, Dr.Sahil Sharma and Dr.Sawinder Kaur for their constant inputs regarding my work, whenever needed.

(Priya Arora)

Abstract

Identifying disease-genes from human genome is a significant and essential issue in biomedical research. Despite several publications using machine learning methods to find new disease genes, it is still difficult due to the factors like pleiotropy of genes, the limited number of confirmed disease genes in the entire genome and the genetic heterogeneity of diseases. Recent approaches have applied the concept of '*guilty by association*' to investigate the association between a disease phenotype and its causative genes, which means that candidate genes with similar characteristics as known disease genes are more likely to be associated with diseases. However, due to the fact that only a small number of genes in the human genome have been experimentally proven to be linked to disease, semi-supervised approaches like positive-unlabeled learning and label propagation are used to find candidate disease genes by training on unknown genes. This is usually the case when there are a small number of confirmed disease genes (labelled data) and a large number of unknown genome regions (unlabeled data). The performance of Disease gene identification models is limited by potential bias of single learning models and incompleteness or noisy biological data sources, therefore ensemble learning models are applied via protein sequences to obtain better predictive performance.

In this work, various machine learning classifiers are analysed and feature extraction method is proposed to choose a more relevant feature set for analysis. An ideal multi-level voting model is proposed, which integrates various ML models based on their False Positive rates to retrieve a new voting classifier for better prediction analysis. The developed model helps to solve the trade-off issue between accuracy and efficiency. A deep learning based methods have also been designed using the Multi-Layer Perceptron (MLP) and Long Short Term Memory (LSTM) for PD genes identification. A comparative study with existing systems shows the effectiveness of the proposed approaches.

Further, disease gene identification is a positive-unlabeled problem. A Positive unlabeled approach have recently been put forth to develop a classification model where known genes are treated as positive training set P and unknown genes are treated as unlabeled set U (instead of negative set N) because unknown genes contain unidentified disease genes. Twelve physicochemical properties of amino acids are applied to generate features with Geary Autocorrelation, Normalized moreau-broto autocorrelation and moran autocorrelation representation methods. The protein sequences based on previous knowledge are adopted to extract features. Consequently, t-SNE is applied to extract relevant features. On the positive unlabelled data a novel n-semble method was proposed which trained a

neural network in a special way and integrated three classification methods based on their F-Score to ensemble the predictions for achieving more accurate predictive analysis. It is found that physicochemical properties of amino acids are highly beneficial in extracting features. Compared with the previous methods on unbalanced datasets, the F Score is improved with proposed n-semble method. The GA representation method characterizes a higher success rate than other representation methods. The experiments were conducted to identify novel disease genes from the entire unlabeled gene set using n-semble algorithm. As a case study, we selected Parkinson's disease category and discovered that several of these identified genes are linked to Parkinson's disease based on the literature survey.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AAC	Amino Acid Composition
AD	Alzheimer's Disease
ANNs	Artificial Neural Networks
AUC	Area Under ROC Curve
Bi-LSTM	Bidirectional Long-Short Term Memory
DNN	Deep Neural Networks
FP	False Positive
FN	False Negative
FPR	False Positive Rate
GO	Gene Ontology
GWAS	Genome-Wide Associations Studies
GA	Geary autocorrelation
KNN	K-Nearest Neighbour
LR	Linear Regression
LSTM	Long Short-Term Memory
ML	Machine Learning
MLP	Multi-Layer Perceptron
MAE	Mean Absolute Error
MSE	Mean Square Error
MA	Moran AutoCorrelation
NA	Normalized moreau-broto AutoCorrelation
PCC	Pearson Correlation Coefficient
PPI	Protein-Protein Interaction
RN	Reliable Negative
ROC	Receiver Operating Characteristic
RF	Random Forest
SGD	Stochastic Gradient Descent
TP	True Positive
t-SNE	t-Distributed Stochastic Neighbor Embedding
TN	True Negative

Chapter 1

Introduction

Complex diseases are caused by the malfunctioning of a group of genes, known as disease-associated genes, or simply disease genes. Identifying these genes is critical for scientists to decipher the mechanism of diseases, which is beneficial to disease diagnosis and drug development [1]. However, this issue is still challenging because identifying these disease genes experimentally is time-consuming and expensive. On the one hand, scientists need to conduct a few experiments to determine whether a gene is disease-associated or not, which may require years of efforts [2]. Moreover, experimental techniques such as Genome-Wide Association Studies (GWAS) are used to identify hundreds of candidates where scientists have to determine the priority of validations to maximize the yield of their experiments. Therefore, computational methods which prioritize disease genes are valuable for disease-gene identification. Currently, many algorithms have been proposed to predict disease genes. Despite their success, different methods have been developed which have their pros and cons. Machine Learning solves the problems by analysing and learning from vast datasets at speeds and capacities not possible for humans alone. Its flexibility in data integration is also valuable for disease-gene prediction because the key for accurate prediction is to properly fuse multi-levels of biological data. Thus, this thesis mainly focuses on machine learning-based methods, which can characterize the non-linear relationships among different types of data.

1.1 Neurological Diseases

Neurological diseases are diseases that disrupt the normal functioning of the nerves, spinal cord, or brain [3]. Abnormalities in the brain structures can result in a variety of symptoms, including altered states of consciousness, paralysis, muscle weakness, poor coordination, seizures, discomfort, and disorientation. These diseases may have a variety of causes, such as environmental factors, infections, genetics, or way of lifestyle choices. Some diseases are present from birth or develop in early childhood, while others can be traced to tumors, trauma, or other structural abnormalities [4]. There are a number of neurological conditions that have direct effects on the brain, which is an essential component of the body and is responsible for the regulation of many important pro-

cesses, including cognition, coordination and movement, emotions, as well as learning, and memory [3]. The hippocampus, which is located in the temporal lobe, is one of the most important areas of the brain. Damage to the hippocampus is linked to a number of neurological and psychiatric illnesses, including Parkinson's disease (PD), Alzheimer's disease (AD), epilepsy, and major depressive disorder, among others [5]. Regardless of the cause, neurological disorders may be linked to temporary or permanent nervous system damage [6]. One such neurological condition is PD, which is largely characterized by cognitive impairments, especially difficulty with memory and learning.

1.1.1 Parkinson's Disease

Parkinson's disease (PD) is a progressive neurodegenerative disease associated with central nervous system affected by the loss of a neuro-transmitter called dopamine. The existing neurons in the brain are responsible for the production of dopamine. The level of dopamine is reduced when the neurons die, which causes the movement problems in Parkinson's [7]. When the level of dopamine decreases, symptoms such as slowness, tremor, and stiffness occur. People with Parkinson's disease have lower dopamine levels than healthy people. Dopamine level in healthy and Parkinson's affected neuron is shown in Figure 1.1. James Parkinson was the first person who announced PD as a "shaking palsy" in 1817 [8]. PD is the second most common neurological disorder in adults after Alzheimer's. This disease mainly affects in Australia, Canada, USA, and other European countries.

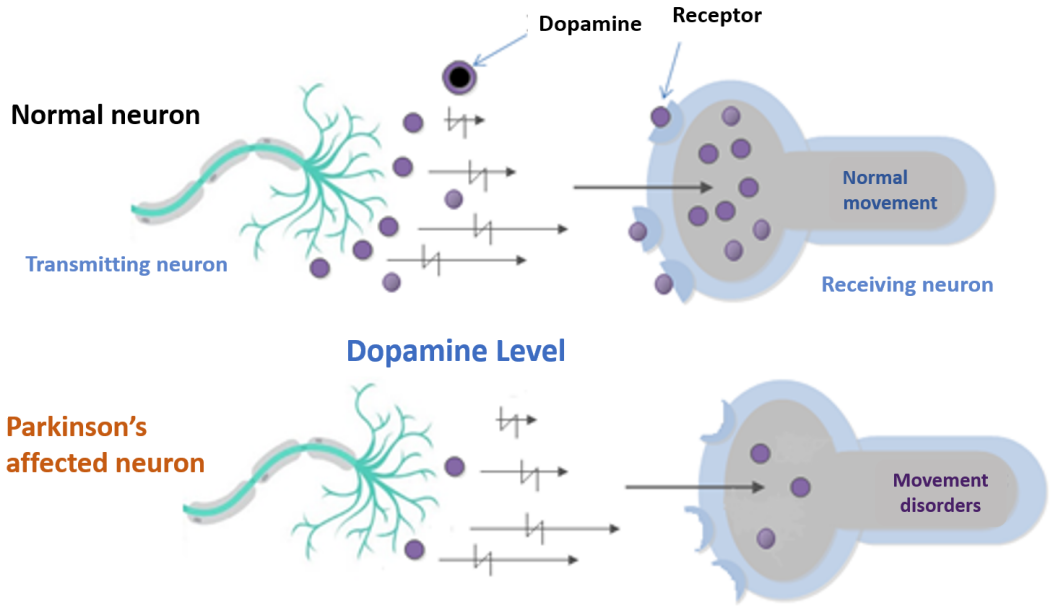


Figure 1.1: Healthy and Parkinson's affected neuron

PD may cause muscle stiffness, slow movement, tremors, and changes in language and writing skills [9]. PD is more common in the elders, with a normal age of beginning of 60 years. The precise reasons for PD are not clear, but studies show that genetic and environmental factors, oxidative stress, and aging play an essential role in the progressive death of dopaminergic neurons.

1.2 Disease Gene Identification

A gene is the fundamental structural and functional component of heredity that controls a variety of biological functions in an organism. The mutation in a single gene sequence may mutate a biological process and leads to a certain disease. A single gene sequence mutation may change a biological process and cause a particular disease. Because human genes interact with one another and are not isolated within the body, a mutation in one gene may have an impact on another gene that interacts with it, which in turn may contribute to the mutation of other genes involved in various biological processes and lead to various disorders. Therefore, taking into account the biological mechanisms and, based on these mechanisms, learning how diseases and genes are related is a significant issue in contemporary biology and medicine.

However high-throughput genomics research led to the discovery of hundreds and thousands of candidate disease genes. The identification of genes involved in specific human diseases has remained a fundamental challenge, requiring time-consuming and expensive experimentation [10]. Computational methods that can reliably predict new disease genes from the huge number of unknown genes will be a good way to speed up the long and hard searches for the genetic causes of different human diseases. Experimentally large number of genes have been identified to be the basic cause of different human diseases over the years. It will be helpful to create machine learning methods that use confirmed disease genes as positive training examples to find new disease genes. This is because genes linked to similar disease symptoms are likely to have similar biological properties. Also, disease phenotype similarity data shows that genes linked to the same or similar disease phenotypes probably have the same biological functions. Several ways have been suggested to rank candidate genes based on different types of biological data, such as evolutionary features, gene sequence data, functional annotation data, PPI (Protein-Protein Interaction) dataset, and gene expression profile.

Disease gene identification is the task of identifying the most plausible candidate disease genes. Disease gene identification is a process by which scientists identify the mutant genotypes responsible for an inherited genetic disorder. Mutations in these genes can

include single nucleotide substitutions, single nucleotide additions/deletions, deletion of the entire gene, and other genetic abnormalities. Understanding the association between casual genes and their genetic disease is a fundamental problem regarding human health [11]. Technology is involved in the detection and monitoring of various human diseases such as Parkinson's. Also, the Internet of Medical Things (IoMT) is in focus for addressing human health. Different experimental methods have been proposed to associate genes with a disease but these methods are expensive in terms of cost and time. For that reason, alternative computational approaches are gaining popularity for disease gene identification.

1.2.1 Motivation and objectives of Disease Gene Identification

In human beings, there are about 2000 monogenic syndromes (syndromes found to be related with a single causative gene). Each syndrome is distinguished from other syndromes by only one or a few of the phenotypic features that it possesses, which are the biological manifestations of the genes that underlie it [12]. Therefore, determining the phenotype-gene relationship is an essential task in biomedicine, which helps biologists and physicians to find the pathogenic process of syndromes. Knowing which genes are responsible for which disease will make patient diagnosis much easier and will provide insight into the functional properties of the mutated gene.

Most of the early methods, including SUSPECTS [13], POCUS [14], ToppGene [15], and ENDEAVOUR [16] have prioritized candidate genes by annotating them with respect to biological structures or functions and comparing their annotations with those of already known disease genes. These annotation-based approaches are limited in that they fail to capture indirect relationships between genes whose common features or functions are not yet annotated. To address the challenge, it is necessary to prioritize candidate genes from hundreds of experimentally suspicious genes using computational techniques, which would greatly reduce the numbers of genes for wet-lab experimental analysis. The common strategy is to rank these candidate genes according to their functional similarity to known disease genes known as disease gene prioritization.

It should be noted that the above methods only provide a gene rank list and a threshold is needed to decide whether a specific gene is disease related or not. A more biologically meaningful approach would build a binary classification model that can automatically identify a gene as disease or not, according to various features of biological datasets, such as protein sequence, PPI and gene expression.

The overall objective of our studies is to use protein sequences (genes) with machine

learning models to improve the accuracy of disease-gene prediction. Before applying machine learning models to this area, a few issues should be addressed. First, considering that we use supervised models to predict disease-gene associations, both positive and negative instances are required to train the models. However, disease-gene prediction is a positive-unlabeled learning problem, in which only positive instances (disease genes) are available. A set of negative instances (non-disease genes) have to be defined before training the models. Thus, developing a strategy to select negative data is fundamental to our research. The next step is to fuse multiple machine learning methods to create an ensemble method for achieving accurate prediction.

1.2.2 Data Sources

Different types of data have been used to predict the disease relevance of candidate genes. To correctly interpret the obtained prediction results, it is necessary to consider the type of evidence used to derive them and to know about possible inherent problems, such as a potential bias towards well-characterized candidate genes. The amount and quality of data used can also have a significant impact on the reliability of the results.

i. **Protein–protein interactions** - The protein interactome, which is the network that shows how proteins physically interact with each other, is one of the most common ways to predict disease genes [17]. This is because it seems obvious that proteins that physically interact with each other often do so to do the same thing. As a result, a negative change to any one of them is likely to result in the emergence of phenotypes that are similar to each other. The widespread correlation between protein complexes and human disease [18] really supports this supposition. Lack of sufficient high-quality experimental data is a serious issue. Many protein-protein interactions are derived from experimental methods like mass spectrometry and the yeast two-hybrid method, which still have issues with sensitivity and specificity.

ii. **Sequence data** - Data obtained through next-generation sequencing techniques with the goal of directly identifying mutations in the genomes of patient's and evaluating their potential disease relevance is a type of evidence that is rarely used but whose importance will certainly increase in the future. This is because next-generation sequencing techniques aim to directly identify mutations in the genomes of the patient's.[19].It is important to highlight the conceptual difference between general properties that genes or proteins show across the entire population (such as their length, degree of conservation, etc.) and case- or patient-specific properties like structural variants and amino acid substitutions.

iii. **Gene expression information** - An essential component of gene function is gene expression. In fact, cellular functions are the result not only of the molecular functions of each part of the cell, but also, to a large extent, of how they work together in space and time. In other words, a gene product's molecular function is largely determined by its enzyme activity, its ability to bind to DNA, or its interactions with other molecules in the cell. However, gene expression is one of the main factors that determine when and where this function is carried out. So, gene expression patterns can tell us a lot about how single genes and groups of genes work together and with each other [20].

    There may be certain limitations even though this information can be regarded as potentially unbiased. For instance, gene expression levels generated from microarray data are unaffected by the degree of understanding of a gene's function so long as the microarray platform has probes or probe sets that target the gene. However, microarrays do not offer a comprehensive coverage, and the design of the platform is biased toward known genes. However, for the development of many platforms, both known and predicted genes were taken into consideration. RNA-Seq is less troublesome, but the quantity of data that is publicly available is not yet equivalent to the enormous number of microarray experiments that have been deposited in public repositories such as the Gene Expression Omnibus (GEO) [21].

iv. **Multiple data sources** - It is rare to predict disease genes from a single form of evidence. As a result of the fact that multiple data sources might provide highly supplementary disease-related information in many instances, they are typically physically and conceptually merged. Even when the transcriptional correlation between two genes is weak, protein–protein interactions can imply functional linkages [16] [22]. A substantial transcriptional coexpression can indicate a functional link between gene products that do not physically interact.

v. **Intrinsic gene properties** -Intrinsic gene or protein properties such as gene or protein length, phylogenetic breadth, degree of conservation, and paralogy – which statistically differ between disease genes and genes not known to be involved in disease – may already provide a hint about a possible relevance for hereditary disorders. This is utilised by a number of prediction tools.

1.3 Research Contributions

The main contribution of this research is to develop a positive unlabeled learning method to identify disease genes from hundreds of candidate genes. The overall contributions are

as follows:

i. A comprehensive literature review on disease gene prioritization methods, machine learning methods which formulates the problem as classification, for neurological diseases have been explored. This review highlights the existing methods, dataset and its features, and comparisons with state-of-the-art methods.

ii. The dataset is formulated by obtaining human Parkinson's (PD) and non-Parkinson's (nPD) protein sequences (genes) from NCBI [23], Ensembl [24] and UNIPROT [25] databases. The extracted sequences are saved as fasta file. The obtained dataset is then cleaned by eliminating duplicate and partial protein sequences.

iii. Various physicochemical properties of amino acids have been used to extract the feature vector.

iv. Resampling techniques have been used in this work to solve the class imbalanced problem as protein sequences (genes) with positive class are large in number as compared to unknown (negative) class. Statistical methods used Friedman and Wilcoxon signed-rank test to find the stochastic nature of algorithms used in the resampling techniques.

v. A two-level ensemble method is proposed with majority voting based on the false positive rate to classify positive and negative samples. Comparison of proposed method is done with deep learning based methods and the state-of-art methods.

vi. A two-level ensemble method is proposed with majority voting based on the false positive rate to classify positive and negative samples. Comparison of proposed method is done with deep learning based methods and the state-of-art methods.

vii. A novel neural network-based ensemble (n-semble) method is proposed focused on positive unlabeled learning classifiers for more accurate and robust disease gene identification. Artificial neural networks (ANN) includes, Multi-Layer Perceptron (MLP) and Long Short-Term Memory (LSTM) are trained and tested to identify Parkinson's disease genes. Finally, a case study is done to predict the novel disease genes of Parkinson's and Cancer diseases.

1.4 Thesis Organization

The thesis is organized into seven chapters. A brief outline of these chapters is given below:

Chapter 1: Chapter 1 begins with a brief introduction of neurological disorders followed

by the problem of disease gene identification. This chapter includes the motivation and challenges of the research. The chapter concludes with the thesis organization, research methodology, and thesis contributions.

Chapter 2: . This chapter provides a comprehensive survey on disease gene identification methods using various biological data sources and computational strategies that combine multiple data sources and learning methods. Various essential research contributions in the field of Intelligent Computing Methods, deep learning and ensemble learning are studied, along with the research gaps, followed by research objectives. The review of disease gene prediction has been performed on the basis of classification problem.

Chapter 3: This chapter provides the overview of techniques that has to be used in the thesis. It underlines a brief introduction of the various intelligent computing methods, deep neural network, ensemble methods, and performance evaluation measures. In this research, we use protein sequences to characterize genes and used representation methods to extract information encoded in proteins. So, different representation methods and physicochemical properties of amino acids are described. This chapter also includes the statistical tests to do statistical analysis during research.

Chapter 4: In this chapter, an ensemble model is proposed to identify Parkinson's disease genes. Various machine learning models are used to build ensemble method using physicochemical properties of amino acids such as Hydrophobicity and Amino Acid Composition (AAC) to classify Parkinson's disease genes. Machine learning approaches, such as Naive Bayes (NB), Support Vector Machine (SVM), K-nearest neighbor (KNN), Neural Network (nnet), Logistic Regression, Decision tree (DT), Random Forest (RF) are effectively used to diagnose PD. Various single classifiers were extensively used for identification of disease genes, but such methods can't get results of an ideal classifier. Therefore, 2-level ensemble method with majority voting based on the false positive rate to develop a classifier with significant improvements compared to existing methods is proposed in this chapter. The methodology followed by the architectural framework of the proposed ensemble model has been visualized to identify disease genes. The motivation behind the proposed model is also discussed in detail in this chapter. The comparison of proposed ensemble method with various existing methods has been also discussed.

Chapter 5: This chapter focuses on the positive unlabeled learning algorithm for disease gene identification where known genes are treated as positive training set P and unknown genes are treated as unlabeled set U (instead of negative set N) because unknown genes contain unidentified disease genes. We have employed twelve physicochemical properties of amino acids to represent the amino acid features. Then dimensionality reduction is done using t-Distributed Stochastic Neighbor Embedding (t-SNE) method. Jaccard

method is applied to find likely negative samples from unknown (candidate) genes. We compared our proposed work with existing state-of-art and ensemble techniques. A case study is done to predict novel candidate genes for Parkinson's and cancer diseases.

Chapter 6: In this chapter, Artificial Neural Network based models including, MLP and LSTM are trained and tested to identify Parkinson's disease genes. The performance of MLP and LSTM methods has been studied on the imbalanced dataset. Firstly, the optimal number of features extracted through feature selection and backward elimination method has been reviewed and optimised. Then, the influence of sequence representation methods has been evaluated on the performance of both MLP and LSTM methods. Finally, a comparison between our proposed method and other disease gene identification methods has been done to obtain relatively negative data to confirm the effectiveness of method.

Chapter 7: This chapter summarizes the key findings of this research and suggests possible directions for research in future.

Chapter 2

Literature Review

In this chapter, we give a comprehensive evaluation of methods for finding disease genes, including computational strategies that integrate several data sources and machine learning techniques. Examining research gaps and significant contributions to machine learning and ensemble learning. Using the identified research gaps as a guide, this chapter summarizes the objectives and innovative contributions of the thesis.

2.1 Neurological Disorders

Neurological disorders, also called brain, behavioural, and cognitive disorders, usually affect a person's ability to walk, talk, learn, and move. These are life-threatening diseases that can have direct effects on the brain and spine [3]. These are the types of conditions that affect people on a chronic basis and are directly linked to the neural system of the human being. Some of the most common neurological diseases are Alzheimer's, Parkinson's, epilepsy, multiple sclerosis, Stroke, autism, and migraines. It is said that age-related diseases like Parkinson's disease (PD) become more common as the elderly people grow [6],[7]. PD, the most common form of dementia, is an irreversible and incurable neurodegenerative disorder associated with a progressive deficiency in memory and cognitive abilities, loss of automatic movements, and problems in daily activities [6],[7], [8]. However, it is believed that genetics have a significant influence in PD risk [7], even though the cause of PD remains unknown. The pathological hallmark of Parkinson's disease [8], [9] is the growth of neurofibrillary tangles and amyloid plaques that can hinder the transmission of information between neurons. This occurs in the loss of nerve cells, which causes the cerebral cortex to atrophy and the ventricles to swell. There are presently no cures or treatments that can arrest the progression of the disease. However, some supportive treatments may temporarily improve the symptoms. This study's major objective was to determine how well machine learning methods can operate as a diagnostic tool for Parkinson's disease. This research was also conducted to test and compare the efficacy of existing ML approaches for the early detection of PD.

2.2 Machine learning-based approaches for disease gene prediction

The disease-gene prediction was often handled as a binary classification issue, with positive and negative training samples comprising the training set. Positive training samples consisted of known disease genes, while negative training samples were frequently selected at random from the remaining ones. Unidentified disease genes may be found in the remaining collection, whereas the non-disease genes utilised in the negative training samples should be genuine. However, there is no database of such genes (no proved negatives) because, in biology, the absence of an association does not necessarily suggest that it does not exist. In order to reduce this uncertainty, one-class classification, in which the classifier is learnt solely from known disease genes, is an alternative method for disease gene prediction. Due to the possibility that the remaining collection contains unidentified disease genes, Semi Supervised Learning (SSL)-based methods were proposed to solve the problem, in which the classifier is learned from both labelled and unlabeled sets.

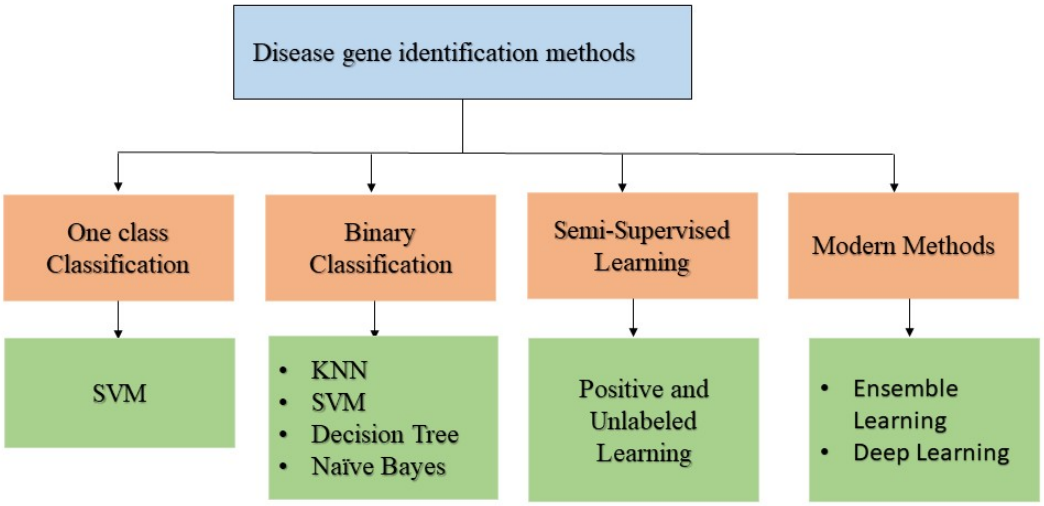


Figure 2.1: The state-of-the-art for disease gene identification classification methods

Positive Unlabeled (PU) learning-based strategies were proposed in particular. The classifier is trained using both positive training examples (i.e., known disease genes) and

unlabeled samples in this method (the remaining genes). Recently, more sophisticated algorithms capable of combining single classifiers (i.e. ensemble learning models) or capable of learning the representation of data (i.e. deep learning models) have been developed for the disease gene prediction problem. Frequently, the performance of a machine learning-based method depends on the quantity of training data (i.e. known disease genes). Figure 2.1 shows the state-of-art for disease gene identification classification methods. In order to improve prediction performance, machine learning-based approaches typically employ all disease genes during training. Table 2.1 summarizes machine learning-based methods proposed for the problem.

2.2.1 One-class classification

The objective of one-class classification is to learn a model that recognizes instances of a specific class, in this case the positive class. Akram *et al.* (2019) [36] proposed a novel One-class Classification Support Vector Machines (OCSVM) technique to accurately classify potential disease genes. The team's objective was to construct a model with enhanced sensitivity and specificity for identifying genes already known to be involved for specific disorders. As a reference disease, they utilized the gene expression data for Acute Myeloid Leukemia (AML) to examine the influence of the proposed model. The data sources for unary and PU learning classifiers are represented as kernel matrices, whereas Duc and Manh (2015) [37] employed vector representations for binary classifiers. Only different types of data can benefit from the kernel-based fusion technique, and a comparison based on the different data representations would be unfair. In this work, they examined multiple classification approaches for identifying disease genes using vectorial representations of samples. The simulation results demonstrated that the unary classification method, which combines density and class probability estimation techniques, achieved the best performance. Abdulaziz and Nasrollah (2015) [19] developed a sequencing-based single-class classification method. Support Vector Data Description (SVDD) is used to train the model and choose important features using PCA. The reason is that, there is no guarantee about the negative instances which are extracted from unknown genes. Therefore, the authors preferred to train the model using only the positive instances. Since the unknown genes set is often comprised with some disease genes, some methods have attempted to reduce the confusion in classification process by selecting a small fraction of unknown genes as negative set [31]. The negative results produced from unlabeled genes are coupled by noisy data, which renders these techniques neither reliable nor robust. One-class classification-based techniques, such as one-class SVM, appear to be the best option, as only a positive training set consisting of known disease genes is required for the training assignment. However, this is especially true for large dimensions. A summarized

Table 2.1: Machine learning-based approaches used for the disease gene prediction

Study	Method used	Dataset	Disease
De Bie T *et al.* [22]	One-class SVM	Biomedical data	General
Abdulaziz *et al.* [19]	Support vector Data Description	Protein sequences	Parkinson's disease
Euan *et al.* [26]	Decision Tree	Protein Sequences	General
Aaron *et al.* [27]	SVM	PPI data, protein sequences	General
Predrag *et al.* [28]	SVM	PPI network, data ontology, protein sequences	General
Shivakumar *et al.* [29]	SVM	PPI network, gene expression	Immunodeficiency disease
Jiabao *et al.* [30]	Artificial neural network and MLP	PPI	Cancer and Diabetes
Peng *et al.* [31]	Positive unlabeled learning (Weighted SVM)	PPI Network, gene ontology	Cardiovascular and endocrine diseases
Fantine and Jean [32]	Positive unlabeled learning	Protein sequences	General
Peng *et al.* [10]	Positive unlabeled learning {Ensemble)	Gene expression, PPI network	Diabetes and cancer
Duc-Hau *et al.* [33]	Random Forest	Biomedical data	General
Ranjan Kumar *et al.* [34]	Deep neural network	PPI network, protein sequence	Alzheimer's disease
Peng *et al.* [11]	Graph convolutional network	Gene expression, disease similarity	General
Jiajie *et al.* [35]	Node2vec and autoencoder	PPI network	Parkinson's disease

review of various state-of-the-art methods for one-class classification of PD patients is presented in Table 2.2.

2.2.2 Binary classification

Traditionally, the prediction of disease genes was addressed as a problem of binary classification, with positive and negative training samples used to train the model. Positive training samples consisted of known disease genes, whereas negative training samples

Table 2.2: Summarized review of research studies related to one-class classification

Author	Find	Method used	Dataset	Features
Akram *et al.* (2019) [36]	The result demonstrates the superiority of the proposed technique to discover disease-causing genes in terms of recall, precision, and F-measure	One-class Support Vector Machines (OCSVMs)	Gene expression	Gene expression omnibus
Abdulaziz *et al.* (2015) [19]	The results demonstrate the significance of resolving the disease identification problem as a single class classification problem	Principal Component Analysis (PCA) and SVDD algorithm	Protein sequences	Physicochemical properties
Duc and Manh (2015) [37]	demonstrated that, compared to one-class SVM based methods, the unary classification technique which combines density and class probability estimation strategy, achieved the best performance	SVM and Random forest	Genomic and proteomic data	Topological properties

were generally selected at random from the remaining genes. The remaining collection may contain previously unidentified disease genes, but the non-disease genes utilised in the negative training samples should be exactly that. However, there is no repository for such genes (no proven negatives). Binary classification requires learning a model that can distinguish between positive and negative examples.

A summarized review of various state-of-art methods for binary classification of PD patients is presented in Table 2.3

Jianzhen and Yongjin (2006) [38] combined topological features of proteins with K-nearest neighbor (KNN) classifier to identify disease-associated genes. They achieved an overall prediction accuracy of 0.76 in a cross-validation test. An SVM-based disease-gene classification method was created by Aaron *et al.* (2007) [27] which takes into account topological aspects of protein interaction networks in addition to sequence-derived and other information. Several novel topological, sequence, and functional traits were discovered that can be used to classify genes for inherited diseases. It was demonstrated that

their approach outperforms the KNN method.

Predrag *et al.* (2008) [28] proposed an algorithm for gene disease association identification detecting based on the protein–protein interaction network, protein sequence, and protein functional information. PhenoPred is a supervised technique, meaning that researchers mapped each gene/protein into the spaces of disease and functional terms using the distance between each gene/protein and all annotated proteins in the protein interaction network. They then used the trained support vector machines to detect gene-disease associations, demonstrating that successful identification of candidate genes is possible even when a large number of candidate disease terms are predicted simultaneously, despite noisy/incomplete experimental data and incomplete disease ontologies.

Keerthikumar *et al.* (2009) [29] identified genes associated with Primary Immunodeficiency Disease (PID) using a two-class SVM classifier. This classifier was trained to differentiate PID genes from other genes using 69 binary features. Using a trained classifier, 1442 probable PID genes were predicted.

Using differential expression data from several diseases, Xiao et al. (2011) *et al.* (2011) [39] constructed ANN classifiers with three layers (i.e. input layer, hidden layer, and output layer), where the number of input neurons was fixed to a specific number of case-control expression datasets. Both the positive and negative training sets included genes known to cause diseases, whereas the negative set was constructed from a random subset of all genes not previously associated with disease.

Abdulaziz and Nasrollah (2013) [9] provided a sequence-based, fast, and customizable PPI prediction method in order to classify the interaction between two proteins (yes or no). Initially, numerous representation strategies have utilised the twelve physicochemical properties of amino acids to convert the sequence of protein pairs into a range of feature vectors in an effort to enhance the presentation of the sequences. Then, principal component analysis (PCA) is executed as a suitable feature extraction method to accelerate the learning process and reduce the influence of noisy PPI data. A novel Learning Vector Quantization (LVQ) predictor has been designed to deal with diverse data models.

Wook (2017) [40] proposed a unique Stepwise Random Forests (SRF) method for feature selection and to improve the identification of disease genes. Two phases comprise the SRF technique. Firstly, the relevant biological features are found using a forward selection method and one-dimensional random forest regression with the updated residual vector as the response vector. Secondly, disease genes are identified using random forest classification in accordance with the selected biological features. Extensive experiments verify the superiority of the proposed SRF method for finding disease-causing genes over

contemporary feature selection and classification methodologies.

Disease-Gene Association (DGA), created by (2018) [41], is a new network-based technique that can determine the score of association between a query gene and a broadened set of disorders. Before calculating the association between two interacting proteins and disease, a large-scale protein interaction network was constructed. Using information about neighboring proteins, the algorithm discovered novel plausible disease genes. The association coverage of genes and diseases was computed and utilised with the association score to perform gene and disease selection in order to discover interesting candidates of disease-gene relationships.

Combining sequencing and protein interaction network data, Ranjan *et al.* (2019) [34] developed a classification method based on machine learning to discover host genes related with infectious illnesses. Deep Neural Networks (DNN) models using 16 characteristics for pseudo-amino acid composition (PAAC) and network attributes had the maximum accuracy 86.33% and sensitivity 85.61%.

Misba *et al.* (2020) [42] presented and analysed new computational algorithms for identifying genes associated with disease. In order to discover new candidate genes, scientists have presented a number of novel topological and biological properties that are currently disregarded. Simulation results demonstrate that the proposed Deep Extreme Learning Machine (DELM) improves upon the state of the art in machine learning accuracy.

2.2.3 Semi-supervised learning

All binary classification-based methods have the same task, which is to find non-disease genes to use as the negative training set for binary classifiers. This is almost impossible to do in biomedical research. To get around this problem, recent studies have suggested that the scoring function can be learned only from known and unknown gene sets. This is called PU learning, which stands for positive (P) and unlabeled (U) learning [44], [45]. This is known to be the most common way to rank a set of data based on how similar it is to a set of good data [46]. PU learning is a set of semi-supervised techniques for training binary classifiers on only positive and unlabeled samples. This machine learning method was used to disease gene prediction problem [31], [32].

Fantine and Jean (2011) [32] introduced a new method (ProDiGe) for prioritisation of disease genes based on the guilt-by-association concept. The ProDiGe method relies on a preexisting set of relationships to infer novel gene-disease links. It allows to combine information from different sources about genes, share information about known disease genes across diseases, and search the whole genome for new disease genes. ProDiGe has

Table 2.3: Summarized review of research studies related to binary classification

Author	Find	Method used	Dataset	Features
Jianzhen and Yongjin (2006) [38]	gained overall prediction accuracy of 0.76 in cross-validation test	KNN classifier	PPI data	Topological properties
Aaron *et al.* (2007) [27]	Their findings revealed the benefits of an integrated method, and they hypothesised that constructing an even larger collection of training characteristics, in conjunction with feature selection approaches, may result in a highly effective tool for disease-gene discovery.	KNN and SVM	PPI data	Topological and sequence based features
Predrag *et al.* (2008) [28]	Using SVM to detect gene-disease associations for multiple Disease Ontology terms, the authors demonstrated that successful identification of candidate genes is possible despite the noise/incompleteness of experimental data and the incompleteness of the ontology of diseases.	an algorithm for detecting gene disease associations namely PhenoPred	PPI network data, protein sequence, and protein functional information	

Xiao *et al.* (2011) [39]	Integrating a large number of disease-specific case-control expression data sets, the authors offer a novel Differential expression pattern (DEP)-based prioritisation approach	ANN	Gene expression data	
Abdulaziz and Nasrollah (2013) [9]	developed a sequence-based, rapid, and adaptive PPI prediction approach to assign proteins to an interaction class	Learning Vector Quantization (LVQ)	PPI data	Topological properties
Abdulaziz and Nasrollah (2015) [19]	Fusion based method is proposed to identify genes using protein sequences	Decision tree and SVM	Protein sequences	Physico-chemical properties
Gholam and Eghbal (2016) [43]	presented percepttron ensemble of positive-unlabeled learning (PEGPUL) technique. The testing results demonstrate PEGPUL's reasonable performance when compared to various common disease gene identification methods.	multilevel support vector machine, k-nearest neighbor and decision tree	online Mendelian inheritance in man (OMIM) database	

been demonstrated to outperform state-of-the-art methods for prioritizing human genes. For the purpose of training numerous classifiers to distinguish P from U, ProDiGe uses random subsets of U. These subsets may contain genes for diseases with no known therapies. Since individual classifiers are faulty, overall classifier performance will deteriorate. Therefore, Peng *et al.* (2012) [31] proposed the positive-unlabeled (PU) learning algorithm PUDI in order to develop a classifier employing P and U. Initially, the universe U is divided into four groups: the strong negative group WN, the moderate negative group NM, the weak negative group WN, and the dependable negative group RN. A multi-level classifier is developed utilising weighted support vector machines, the four training sets, plus the positive training set P, to identify disease genes. The experimental findings demonstrate that the proposed PUDI method outperforms the current best methods. By utilising information on both disease genes and disease gene neighbours through a protein-protein interaction network, Thanh and Tu (2012) [47] proposed a novel method for predicting disease genes using the semi-supervised learning (SSL) algorithm. Six distinct biological databases, including Gene Ontology, Pfam, Universal Protein Resource, Reactome, Interologous Interaction Database, and InterDom, along with a gene expression dataset, were merged to give a comprehensive picture of proteomics and genomics. SSL surpasses k-nearest neighbours and support vector machines in terms of sensitivity, specificity, precision, accuracy, and balanced F-function using 10 fold cross validation. Semi-supervised learning made it easier to study disease genes, especially for a specific condition when there were very few known disease genes (labelled data).

For the objective of disease gene identification, Peng et al. (2014) [10] developed a powerful PU learning framework that combines numerous biological data sources with a collection of potent machine learning classifiers. This training technique for PU learning classifiers uses data from multiple biological sources. To increase the accuracy and robustness of disease gene predictions, EPU, a new ensemble-based PU learning method, is used to combine various PU learning classifiers. Experiments comparing EPU to other state-of-the-art prediction methods and ensemble learning classifiers revealed that EPU performed significantly better across six sickness classifications. By mixing the outputs of an ensemble of PU learning classifiers for prediction with different biological data sources for training, they were able to reduce the likelihood of bias and error in the data sources and machine learning techniques. This led to more accurate and robust disease gene predictions.

Abdulaziz and Nasrollah (2015) [48] proposed a novel Sequence-based fusion (SFM) technique for identifying the disease genes. Here, the amino acid sequence of the proteins, which is universal data, is employed to present the genes (proteins) in four unique fea-

ture vectors, as opposed to the noisy and incomplete prior information used by earlier methods. The intersection set of our negative sets, derived using a distance technique, is utilised to suggest potential genes with more probable negative data. After applying independent state-of-the-art support vector machine (SVM)-based predictors, they combined their results using decision tree (C4.5). A summarized review of various state-of-art methods for semi supervised learning of PD patients is presented in Table 2.4.

2.3 Deep learning-based approaches for disease gene prediction

As an alternative to conventional machine learning, researchers are beginning to investigate deep learning models for disease diagnosis. In recent years, Deep Learning (DL) techniques have been successfully applied to disease diagnosis. The Artificial Neural Network (ANN) is the basis upon which DL methods are constructed. These techniques, which can successfully identify latent patterns in the data without explicitly extracting features, are also known as representation learning techniques.

Protein sequence information is used to characterise genes, disease symptoms are used to characterise disease, the combination of protein sequence information and disease symptoms is used to map disease-gene associations into two-dimensional images, and a convolutional neural network is used to establish a predictive model for predicting disease genes Xingyu *et al.* (2020) [49] . In contrast to the training set's 92.38% and 91.17% accuracy and sensitivity, the test set only achieves 80.64% and 80.69% accuracy and sensitivity.

MISBA *et al.* (2020) [42] assessed the performance of various computational techniques on disease-gene association data from DisGeNET using TP rate, FP rate, precision, recall, F-measure, and ROC curve evaluation criteria. The results demonstrate that many computational algorithms employing an enhanced feature set can achieve greater accuracy (up to 93.8%), recall (up to 93.1%), and F-measure (up to 92.9%) than the prior state-of-the-art methods. Thalassemia, diabetes, malaria, and asthma were the four primary conditions they investigated with their methods. In simulated experiments, the proposed Deep Extreme Learning Machine (DELM) delivered more accurate results than previously published approaches.

Ritu and Manik (2020) [50] provide a comprehensive review of deep learning techniques used in the prognosis of eight distinct neuropsychiatric and neurological conditions, including Alzheimer's, stroke, epilepsy, autism, Parkinson's, migraine, multiple sclerosis,

Table 2.4: Summarized review of research studies related to semi-supervised learning

Author	Find	Method used	Dataset	Features
Fantine and Jean (2011) [32]	ProDiGe technique introduces a novel machine learning paradigm for gene prioritization, which could be beneficial for discovery of novel disease genes	SVM	Protein sequences	PPI
Peng *et al.* (2012) [10]	proposed a novel positive-unlabeled (PU) learning algorithm PUDI (PU learning for disease gene identification) to build a classifier using P and U	Weighted SVM	PPI Network, gene ontology	Topological properties
Thanh and Tu (2012) [47]	presented a novel method to effectively predict disease genes by exploiting, in the semi-supervised learning (SSL) scheme, data regarding both disease genes and disease gene neighbours via protein–protein interaction network	KNN and SVM	Universal Protein Resource, Interologous Interaction Database, Reactome, Gene Ontology, Pfam, and InterDom, and a gene expression	PPI
Abdulaziz and Nasrollah (2013) [48]	Introduced a method for predicting protein-protein interactions from protein sequences. Results indicate that neural networks are a promising alternative to many conventional classification methods	Learning Vector Quantization (LVQ) neural network and PCA	Protein sequences	Physico-chemical properties

and cerebral palsy. These diseases are severe, life-threatening, and, in the majority of cases, can lead to other dangerous human diseases. Deep learning techniques are an emerging soft computing technique that has been used profitably to solve a variety of real-world issues, including pattern recognition (Face, Emotion, and Speech), traffic management, drug discovery, disease diagnosis, and network intrusion detection. This study imparts the discipline, frameworks, and methodologies utilized by various deep learning techniques to diagnose various neurological disorders in humans. Ping *et al.* (2019) [51] proposed a method to predict disease–gene relationships using multi-modal DBN (dgMDL). In the beginning, two DBNs independently discover latent representations of protein-protein interaction networks and gene ontology concepts. The combined latent representations of the two sub-models serve as the multimodal input for a joint DBN designed to learn cross-modality representations. The acquired cross-modality representations are then utilised to predict disease-gene relationships. On a curated set of disease-gene associations, five-fold cross-validation is utilised to compare the proposed method to two state-of-the-art methodologies. This capacity of dgMDL to predict novel disease-gene relationships is further illustrated by an analysis of the top 10 unknown disease-gene pairings. Mohamad *et al.* (2019) [52] develop a deep learning technique that combines spatial and sequential properties To predict disease-associated mutations of the metal-binding sites. In this paper, the authors demonstrate that the MCCNN (Multichannel Convolutional Neural Network) can be trained to predict atomic and ionic changes to metals in metal proteins. PADIDEH *et al.* (2017) [53] developed a deep learning technique for detecting cancer and identifying crucial genes for breast cancer diagnosis. They began by employing Stacked Denoising Autoencoder (SDAE) to extract all functional information from high-dimensional gene expression patterns. Next, scientists evaluated the performance of the extracted representation using supervised classification algorithms to confirm the value of the new characteristics for cancer identification. Lastly, they analysed the SDAE connection matrices to identify a group of genes with extensive interactions. On the basis of the data and analysis, it appears that these highly interacting genes may serve as useful cancer biomarkers for the detection of breast cancer; nevertheless, further investigation is required.

Jiajie *et al.* (2019) [35] utilized the Node2vec tool to generate a vector representation of each gene in a PPI network, followed by an autoencoder to minimize the dimension of the resulting vector. Ultimately, new genes associated with Parkinson's disease were predicted using a support vector machine classifier. In addition, they utilised N2A-SVM trained on the most recent dataset to predict genes for Parkinson's disease. In some studies, CNN has been used the similar purposes for physiological affect detection [54] [55] and also used for time series analysis for affect computing [56]. Recently, Deep

learning methods have also been responsible for a variety of healthcare problems, such as disease prediction, image segmentation, audio recognition, medical imaging, etc [57], [58].

A summarized review of various state-of-art methods for deep learning based methods for classification of PD patients is presented in Table 2.5.

2.4 Ensemble learning

Due to the dependence of a single classifier's performance on the benchmark dataset, recent research has integrated single classifiers with ensemble learning approaches to improve disease-gene association prediction. Ensemble strategies have also been utilized to integrate PU learning methods for the problem of disease gene prediction. Ying *et al.* (2017) [59] Due to the dependence of a single classifier's performance on the benchmark dataset, recent research has integrated single classifiers by ensemble learning approaches to improve disease-gene association prediction. Ensemble strategies have also been utilized to incorporate PU learning methods for the problem of disease gene prediction. Gholam and Eghbal (2016) [43] reported the Perceptron Ensemble of Graph-based Positive-Unlabeled Learning (PEGPUL) on protein domains, gene ontologies, and protein-protein interaction networks. This method extracts a set of positive and negative genes using a co-training framework. A Perceptron ensemble is ultimately learned from three weighted classifiers: a multilevel support vector machine, a k-nearest neighbor classifier, and a decision tree. Table 2.6 summarizes the review of various state-of-art methods for the hybrid classification of PD patients.

2.5 Research Gaps

The research gaps have been formulated based on the observations made during the literature survey and are stated as follows:

1. The early stage detection of genomic-based diseases is still a challenging task and also the prediction based on the physical, cognitive and behavioral change of the patient can be considered a major gap in the research.

2. Predictions of diseases based on genomic data have been made using a wide variety of conventional methodologies. These current approaches rank genes as candidates for disease, and a threshold is required to determine whether or not a given gene is causal of the disease.

Table 2.5: Summarized review of research studies related to deep learning

Author	Approach	Technique used	Remarks
Xingyu *et al.* (2020) [49]	identifying potential disease-associated genes	Convolutional neural network	explore the relationships between diseases and genes, and has an important impact on the disease etiology research and pharmaceutical industry.
Ritu and Manik (2020) [50]	Highlight the research work on early diagnosis of neurological diseases using deep learning techniques	Deep Neural Network (DNN), Deep-Belief Network (DBN), Deep Autoencoder (DA) and Convolutional-Neural Network (CNN).	For Alzheimer's and Parkinson's disorder, CNN found to present better results than other DL methods
MISBA *et al.* (2020) [42]	Improve the performance of computational approaches	Deep extreme learning machine (DELM)	
PADIDEH *et al.* (2017) [53]	The results and analysis illustrate that these highly interactive genes could be useful cancer biomarkers for the detection of breast cancer that deserve further studies.	Stacked Denoising Autoencoder (SDAE)	
Mohamad *et al.* (2019) [52]	develop a deep learning approach by incorporating both spatial and sequential features to predict disease associated mutation of the metal-binding sites	Multichannel Convolutional Neural Network	Moreover, the approach can be further exploited for other applications such as prediction of binding affinity of small molecules to proteins, which may guide development of new drugs.
Ping *et al.* (2019) [51]	Proposed a method to predict disease-gene associations with cross-modality features obtained by multimodal deep learning.	Multimodal Deep Learning	Further analysis of the top-10 unknown disease–gene pairs also demonstrates the ability of the proposed model

MISBA *et al.* (2020) [42]	Improve the performance of computational approaches	Deep Extreme Learning Machine (DELM)	
Jiajie *et al.* (2019) [35]	Used N2A-SVM algorithm to discover new genes associated with PD	Node2vec Autoencoder and SVM (N2A-SVM)	Trained N2A-SVM on the recent dataset and used it to predict Parkinson's disease genes.

Table 2.6: Summarized review of research studies related to hybrid classification

Author	Find	Method used	Dataset	Features
Ying *et al.* (2017) [59]	Comparatively analyzed the topological properties between disease-related genes and non-disease genes in protein-protein interaction network	Random forest	PPI	topological features
Peng *et al.* (2014) [10]	Through integrating multiple biological data sources for training and the outputs of an ensemble of PU learning classifiers for prediction, they were able to minimize the potential bias and errors in individual data sources and machine learning algorithms to achieve more accurate and robust disease gene predictions.	Nearest Neighbor, Naïve Bayes and SVM	Gene expression, Gene ontology and PPI network	Biological features

3. There are many experimental approaches that have been introduced in recent years to identify disease genes from vast number of candidate genes. These techniques differ in the genomic data type used to generate feature vectors, such as PPI, gene expression profiles, protein and biological functions. Unfortunately, all these techniques are based on the information of proteins achieved from protein domains, gene ontology, and, PPI data. Hence, might not be implemented accurately as information is incomplete, noisy, and time consuming.

4. Disease gene prediction is typically viewed as a classification challenge using machine learning-based methods. Since there are no established non-disease genes,

identifying the negative training set using a binary classification approach is challenging.

5. It is difficult for a classification model based on a single hypothesis to achieve competitive performance in all disease groups.

2.6 Objectives

To address the aforementioned research gaps, the following objectives were identified for the research work:

1. To explore and analyze existing techniques for the prediction and analysis of neurological diseases.
2. To fetch, pre-process and analyze neurological diseases dataset.
3. To develop an integrated Intelligent Computing framework for the prediction of neurological diseases.
4. To verify and validate the proposed framework with existing methods.

Chapter 3

Background

This chapter discusses the primary concepts and methods that have to be used in the thesis. It underlines a brief introduction to the various machine learning methods, deep neural networks, ensemble methods, and performance evaluation measures. In this research, we use protein sequences to characterize genes and used representation methods to extract information encoded in proteins. So, different representation methods and physicochemical properties of amino acids are described. This chapter also includes the statistical tests to do statistical analysis during research.

3.1 Machine Learning

Machine Learning (ML) is defined as the "Scientific study of algorithms and statistical models that computer systems use to progressively improve their performance on the specific assigned task. Machine learning is the study of learning frameworks that incorporate principles and strategies from the fields of both statistics and mathematics. The area addresses the creation and implementation of intelligent systems capable of learning from data without being directly programmed. The models are used to uncover secret patterns and developments in the data that contribute to meaningful observations and are helpful in making data-driven decisions [18]. As the model learns from data and is able to execute activities from that data, the consistency and quantity of data accessible can dictate how much the model will learn. Machine Learning approaches are frequently related to data processing, predictive data analysis, and computer science processes, but on the actual ground, these concepts vary in nature.

Machine Learning has been widely used for the development of intelligent software for speech recognition, computer vision, robot control, natural language processing, and other applications. Machine Learning is gaining popularity in studying chronic diseases with applications ranging from early prevention, diagnosis to predicting treatment effect and prognosis. In medical sciences today, diagnosis of the disease is a serious task that relies on clinical examination and assessment. Thus, for cost-effective management as well as decision making decision support systems based on computers may play a pivotal

role. The healthcare field creates a huge amount of data which comprises assessment reports including patient's clinical and physical assessments, treatment, future appointments, and a list of prescribed or non-prescribed medicines [60]. Figure 3.1 illustrates the machine learning application in healthcare. It is a tedious and complicated task to manage this data in a required manner so that it can be effectively extracted and efficiently processed.

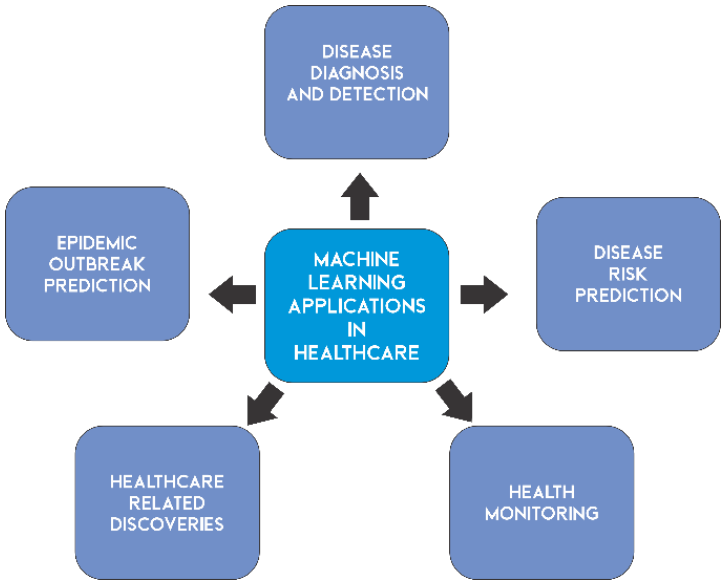


Figure 3.1: Machine learning applications

The major challenges for the healthcare professionals comprise of conditions underlying diseases that may not be directly observable or measurable as data, in finding latent variables is one of the challenges. Thus, Machine Learning algorithms play a critical role in the early detection of diseases. However, other challenges could be that the disease is continuously evolving which would vary from person to person, at times the patient's information is incomplete, and there could be observational differences apart from integrating the domain knowledge which is an essential step in the modelling process. Machine learning techniques are broadly classified as unsupervised or supervised learning techniques depending on the learning process of the model. Classification and regression are supervised methods where the labelled instances are used to train the algorithm, and the outcomes are expected from the model. However, clustering and association analysis are unsupervised learning methods in which the unlabelled data is used by the learning algorithm to prepare the model. Supervised approaches involve classification and regression analysis, while unsupervised strategies include clustering and association analysis.

3.2 Machine Learning methods used

Eight machine learning methods, namely Decision tree (CART), Naive Bayes (NB), Neural Network (nnet, pcannet), Support Vector Machine (SVM), K-nearest neighbor (KNN), Linear Model (Logistic Regression), Boosting method (Adaboost and Gradient Boosting) and Random Forest (RF) have been applied for diagnosis of Parkinson's disease. Table 3.1 shows the summary of machine learning methods with their tuning parameters.

3.2.1 Decision Tree

A decision tree (DT) is one of the supervised learning algorithms, which is used to solve classification problems. It can be used for discrete and continuous input and output variables. While constructing the decision tree top-down approach is considered. Each branch node in the decision tree shows a preferred attribute from the given attributes and each leaf node shows a decision or output [61].

Classification and Regression Trees (CART) CART is a non-parametric machine learning algorithm for building a decision tree mostly used for classification problems. It frequently divides the data and makes homogeneous groups. The CART method involves two steps in building classification trees. Firstly, the root node will split into two child nodes based on the outcome variable's value and then the process of tree pruning is performed in the next step. Pruning helps to minimize loss function [62]. This method can handle both continuous and categorical data to give easy-to-compute and easy-to-interpret estimators. CART is flexible in practice because it makes it easy to model non-smooth or non-linear relationships [63]. CART performance is not affected by outliers of input parameters but imbalanced classes may result in under-fitted trees [64]. The growing stage segments the training examples recursively until the Gini index is minimized at each leaf node. The selection of the best tree is done by pruning method when Eq. 3.1 is minimised for all leaf nodes [65].

$$a(t) = \sum_{a \neq b} \rho(b|t)\rho(a|t) \tag{3.1}$$

where a and b are class labels and p($b|t$) is the conditional probability to detect a sample from class b at node t.

3.2.2 Naïve Bayes (NB)

Naïve Bayes is the simplest probabilistic model among Bayesian networks, used to predict the probabilities of membership for each class [66]. The term Naïve refers to the supposition that different attributes are independent of each other in the given course of instruction. Therefore, NB is defined as the conditional independence probabilistically. Given a set of S protein sequences $S_i = s_1, s_2, \ldots\ldots, s_n$ where every sequence consists of A amino acids, $S_i = a_1, a_2, \ldots, a_m$. Then, the probability of each S_i occurrence in class label C_l is given by Eq. 3.2.

$$P(C_l|S_i) = P(C_l) \prod_{n=1}^{m} P(s_n|C_l) \tag{3.2}$$

Here, the conditional probability of amino acid a_m present in a protein sequence of class label C_l and the prior probability of sequence occuring in class label C_l is denoted by $P(C_l)$. Naïve Bayes is simple to use and can easily handle missing values for these attributes by neglecting the conditional probabilities. It requires only a small amount of training data for probability generation [67]. Therefore, this method is suitable for the proposed method as it contains few Parkinson's disease sequences for training.

3.2.3 Neural Networks

Neural networks are constructed as a group of mutually dependent similar neurons. Interconnects are used to transmit signals from a neuron to another neuron. In addition, these connections may increase neuronal delivery [63]. These models can learn from past experience and to classify nonlinear separable patterns [68]. In this paper, we have evaluated our dataset on Multi-Layer Perceptron (MLP). MLP can be trained on both classification and regression dataset. The feature vector Vf = $f1, f2 \ldots fn$ are retrieved from the feature extraction phase and then the classifier learns a given function in Eq. 3.3 through a training dataset.

$$g(n) : S_i \rightarrow S_o \tag{3.3}$$

where i denote the input dimensions and o denotes the output dimensions. S_i represents the input space or the set of all possible input values that can be provided to the neural network. and S_o is output space could also be a set of numerical or binary values that represents the output produced by the neural network. There can be one or more hidden layers between input and output layers. The input layer is composed of neurons, where

each neuron represents the input feature which is fed into the hidden layer. Then, the hidden layer calculates the weighted summation w1f1+w2f2+... wnfn, followed by function g (n). The output value is given by the last hidden layer and is received by the output layer.

Neural Network with Principal component (PCANNET) is a wrapper function for both preProcess and nnet functions, which will perform a Principal Component Analysis (PCA) on the features before giving input to the neural network. This function will retain sufficient components that are used to capture some pre-defined threshold on the cumulative proportion of variance. Then, the same transformation is applied to the new predictor values for new samples. Pcannet is available for both classification and regression.

3.2.4 Support Vector Machine (SVM)

An optimum edge classifier called SVM offers a statistical learning approach. SVM is a computational algorithm that can learn binary classification problems from training examples [69]. SVM can be defined in four primary standards which includes the separation of hyper planes, the maximum boundary hyperplane, soft boundaries, and kernel functions. The data occurrences are represented as n-dimensional vectors for a normal linear classifier and (n – 1) dimensional vector is used to cut in two positive and negative occurrences. However, the use of kernel function is to determine the gap between data points for non-linear classifiers [70]. Through the SVM model, features retrieved from hydrophobicity and amino acid composition of protein sequences are represented as points in feature space. Then, the features are mapped in such a way that a wide gap is visible to perform linear classification. In our dataset, the features are labeled by making two categories, C= PD, nPD and then the training classifier builds a model which assigns new features to both defined categories.

3.2.5 K-Nearest Neigbhor (KNN)

KNN is a non-parametric classifier used to classify unknown instances represented by certain feature vectors as a point in feature space, which measures the distance between two data points in the training data set. Euclidean distance has been used in this paper as a distance metric for classifying nearest neighbor [71]. k-NN is depicted as an instance based knowledge. Rather than separating rules from the training data, the distance matrix is used to compare new samples with existing ones. Most of the k nearest neighbors is used to predict classes in new cases. The values of k and distance metrics used are the only two adjustable parameters in KNN. k is the number of nearest neighbors to be

included in the class membership estimation: the value of P(y|x) is calculated simply as the ratio of class y members over k nearest neighbors of x. By varying the values of k, we can make the model more or less flexible.

3.2.6 Linear Model

Linear models help in classifying groups by linearly combining feature vectors. If v is the input feature vector to the classifier, the resulting score is given by Eq. 3.4.

$$d = f(\Sigma_i v_i w_i) \tag{3.4}$$

here w denotes feature vector's weight and the function f provides the output of two vectors.

Extending Linear model, Logistic Regression (LR) finds the relationship between dependent and one or more independent variables by using a logistic function to estimate the probability [72]. The independent variables are considered here as predictors of dependent variables and when the dependent variable is in binary form, it can be measured by ordinal, nominal or interval scales. There is a non-linear relationship between the dependent and independent variables. The relationship between occurrence and its dependence on several variables can be represented by Eq. 3.5:

$$p = \frac{1}{1 + e^{-z}} \tag{3.5}$$

Where p is probability of having PD, z is the input and probability can take values from 0 to 1 [73].

3.2.7 Boosting Methods

These methods help to improve the model predictions of a learning algorithm by combining the weak classifiers to obtain a strong classifier. In this research, we have applied Adaboost, Xgboost and Gradient Boosting (GBM) methods.

Adaptive Boosting (ADA): It is one of the most commonly used boosting methods for improving the accuracy of an algorithm by creating a strong classifier with the combination of weak classifiers. This is an iterative method and generates a robust classifier that includes a series of complementary weighted classifiers [74]. The basic idea of this technique is to maintain a distribution (set of weights) on the training set. The weight of each step is initialized to the same value (equal to $1/N$, where N is the number of data

points in the training set). Then, the error (e) is calculated as the sum of the weights of the unclassified points at every step [75]. Weights for correct examples are left as it is and the weights for incorrect examples are updated by being multiplied with $\alpha = (1-e)/e$, and for the sum to be 1 the whole set is normalized. The classification of a new occurrence each having a weight α_t is made by voting on all classifiers. Mathematically, it can be represented as shown in Eq. 3.6:

$$f(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \tag{3.6}$$

where $h_t(x)$ is a weak learner and α_t is corresponding weight [76]. It is exactly the weighted combination of T weak classifiers. $\alpha_t = 0.5 * ln((1-e)/e)$: weight of classifier is based on error rate.

Gradiant Boosting method (GBM): GBM works same as Adaboost by consecutively adding predictors to an ensemble, and each predictor corrects its predecessor. Ensemble model in GBM is also a weighted sum of weak classifiers. The major difference with Adaboost is in the definition of sequential optimisation process [77]. This method tries to fit a weak classifier to the opposite of the gradient of the residual error with respect to the current ensemble model. It is an iterative approach and to find an optimal model in this form is difficult too.

Extreme Gradiant Boosting method (XGBoost):XGBoost is a specific implementation of the Gradient Boosting method which uses more accurate approximations to find the best tree model.It is an efficient and scalable implementation of gradient boosting framework by [78]. XGBoost, a type of gradient boosting, has two major improvements: (a) speeding up the tree construction and (b) proposing a new distributed algorithm for tree searching. XGBoost can automatically obtain the importance of features such that features can be fltered. Therefore, we used XGBoost to select the most relevant features during the initial feature selection [79].

3.2.8 Random Forest(RF)

Random Forest integrates collection of decision trees with ensemble technique. The procedure of tree building is the same as CART strategy but without pruning. Given a dataset D = (xi,yi), i=1,2....N for training where the input matrix X contains N number of samples and the output Y is label of X. To extract the ntree sample set Sk, where k = 1, 2...ntree from the original sample set S, RF method uses the bootstrap resampling technique [80]. The main idea behind RF is that only m features from k features

Table 3.1: Summary of used machine learning methods

Model	Method	Tuning Parameter
Decision Trees (CART)	rpart	Cp : 0.0275, maxdepth, riterion : Gini
Naïve Bayes	nb	Usekernel : TRUE, adjust : 1, fL: 0
Stochastic Gradient Boosting	gbm	None
Adaboost	ada	nu : 0.1, iter : 100, maxdepth : 1
Random Forest	rf	mtry : 2, ntree : 500
Neural Network	nnet	Number of hidden layers :4
PCANNET	pcannet	
Support Vector Machine	svm	Kernel="Radial", type="C-svc"
K Nearest Neighbors	Knn	k=19
Logistic Regression	Glm	None

are considered at each split and these features are chosen randomly [81]. Therefore, the averaging (k-m)/k splits do not consider the strong predictors, thus providing higher probability to other predictors. This method uses bagging to increase the tree diversity by growing trees from different training datasets, thereby reducing the overall variance of the model. It turns out that when the number of trees is large enough, the upper bound of the generalized error coincides to the following formula:

generalization error $\leq \frac{\bar{\rho}(1-s^2)}{s^2}$

where s is a number to measure the strength of tree classifiers and $\bar{p}$ is the average correlation among trees.

3.3 Artificial Neural Network (ANN)

Artificial Neural Networks (ANNs) mimic the human brain through a set of algorithms. Human brains are made up of connected networks of neurons. ANNs seek to simulate these networks and get computers to act like interconnected brain cells, so that they can learn and make decisions in a more humanlike manner. Different parts of the human brain are responsible for processing different pieces of information, and these parts of the brain are arranged hierarchically, or in layers. An ANN can have only three layers of neurons: the input layer (where the data enters the system), the hidden layer (where the information is processed) and the output layer (where the system decides what to do based on the data). But ANNs can get much more complex than that, and include multiple hidden layers [82], [83]. Artificial neural network methods utilize a set of computational layers designed to learn patterns from input data. Each layer is employed to extract a specific type of information. The output of a certain layer is input to the succedding layer. The input data is fed into the input layer and the target is generated by the output

layer. We have designed and compared three ANN models i.e. Multi-Layer Perceptron (MLP) and variants of RNN such as LSTM and Bidirectional LSTM to identify genes responsible for PD.

3.3.1 Multi-Layer Perceptron (MLP)

Multilayer Perceptron is also known as deep feedforward network with multiple hidden layers. Large number of layers enhances the prediction performance of nonlinear tasks [31]. The neurons should be organized in a unidirectional pattern in MLP, where each layer output's forms the next layer inputs. Input layer, hidden layer and output layer are the three types of layers to convert data in MLP. Figure 3.2 depicts the basic architecture of MLP network with two hidden layers. The collection of neurons involves the neural network structure, where each neuron is related to entirely every neuron in the next layer. The connections between these layers should be categorized by certain weights located inside [-1, 1] [84]. Each node in MLP can do summation and activation functions. All the three layers of MLP are associated with the parameters of the MLP network through the weight (w) and bias (b) using summation function as shown in Eq. 3.7. Then multiplied each neuron by weight in the input layer.

$$S_k = \Sigma_{i=1}^{n} w_{ik} I_i + B_k \tag{3.7}$$

where I_i denotes input variable, n denotes number of input, B_k denotes bias and w_{ik} denotes connection weight. An activation function should be activated using the output of Eq. 3.7. The most commonly applied activation functions in MLP are hyperbolic tangent (tanh), sigmoid, Rectifier Linear Unit (ReLU), sigmoid, Leaky ReLU. In this paper, we applied ReLU activation function at hidden layer and sigmoid at output layer as shown in Eq. 3.8 and Eq. 3.9 respectively. Relu is used as default choice for hidden layers, derivative is not 0 (mostly $z > 0$ so derivative is +ve) which makes the learning faster but other activation functions can have derivative 0 which slows down the learning of neural networks.

$$f_k(x) = max(0, S_k) \tag{3.8}$$

$$f_k(x) = \frac{1}{1 + e^{-s_k}} \tag{3.9}$$

Therefore, the final neuron output can be obtained from Eq. 3.10.

$$y_i = f_k(\Sigma_{i=1}^{n} w_{ik} I_i + B_k) \tag{3.10}$$

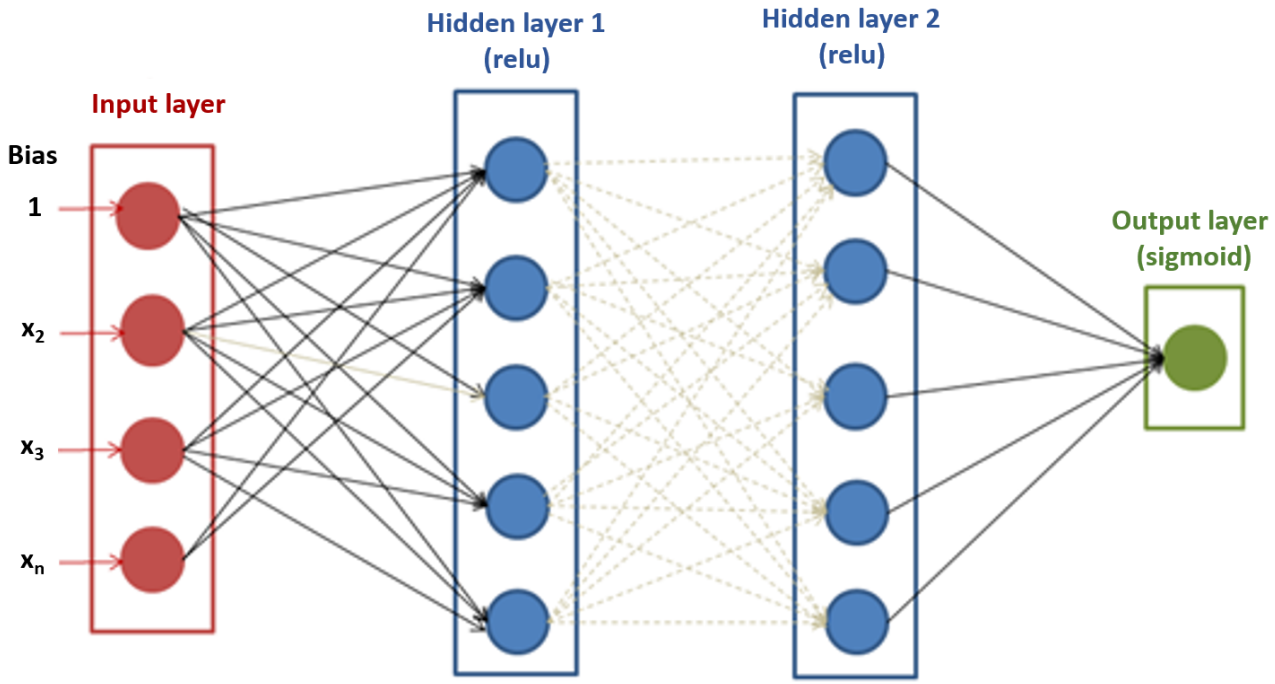


Figure 3.2: Architecture of MLP

3.3.2 Long Short-term memory (LSTM)

LSTM is a special kind of RNN (Recurrent Neural Network), currently become popular in the area of machine learning, introduced by Hochreiter and Schmidhuber [85]. LSTM is designed to learn long-term dependencies and memorize information for a long time. It is organized in the chain-like structure and consists of repeating module as in standard RNN. LSTM is based on three added gates; Input gate, forget gate and output gate as shown in Figure 3.3. LSTM has four layers to interact instead of a single neural network layer. LSTM method comprises of storage blocks called memory cells. The two states i.e. hidden state and cell state are moving to the next cell. Initially, the cell state allows the data to flow forward substationally unchanged. However, certain linear transformations can take place. With sigmoid gates, data can be added to or deleted from the cell state [79]. LSTM aims to avoid long-term dependence issue to control the memory process with the use of gates.

The initial and main step in building an LSTM system is to find and then remove unwanted information from the cell. A sigmoid function helps to identify and excludes unwanted data from cell state, which gets the output of previous timestamp (h_{t-1}) at t-1 time and current input (x_t) at t time with bias b_f as shown in Eq. 3.11. In addition, the sigmoid function decides which part of the previous output could be excluded. The gate is also known as forget gate. Output of sigmoid layer determines whether to completely

retain or completely discard information (0 or 1).

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f) \tag{3.11}$$

where, σ denotes sigmoid function, W_f and b_f are weight matrices and bias respectively.

The next step is to determine and store information in cell state from the new input x_t state. This can be done with sigmoid and tanh function. Sigmoid layer decides whether to update or ignore new information (0 or 1) and tanh layer assigns weight to passed value to determine its importance level (-1 or 1). Then multiply these values to update new cell state and add this new memory (N_t) to old memory $c_{(t-1)}$ resulting in c_t.

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \tag{3.12}$$

$$N_t = tanh(W_n.[h_{t-1}, x_t] + b_n) \tag{3.13}$$

$$c_t = f_t * c_{t-1} + i_t * N_t \tag{3.14}$$

where, c_t and $c_{(t-1)}$ are cell states at t and t-1 time respectively.

The last step determines which cell state information is used as output. This step separates the final memory from the hidden state. As shown in Eq.3.15 sigmoid layer determines which part of cell state makes it an output O_t with last hidden state $h_{(t-1)}$. Then, output of sigmoid function is multiplied with new values formed by tanh function from estimated cell state as shown in Eq.16.

$$O_t = \sigma(W_o.[ht-1, x_t] + b_o \tag{3.15}$$

$$h_t = o_t * tanh(c_t) \tag{3.16}$$

3.3.3 Bidirectional Long Short-term memory (Bi-LSTM)

Bidirectional-LSTM, another variant of RNN developed by Schuster and Paliwal [86] extracts the complete temporal information of time t by considering both past and future information. Recurrent Neural Networks process the inputs in only one direction and ignores the information processed in future. This issue is overcome by following the bidirectional topology of LSTM. In Bi-LSTM, hidden neurons of standard RNN are split into forward and backward states in which neuron of forward states are not connected to

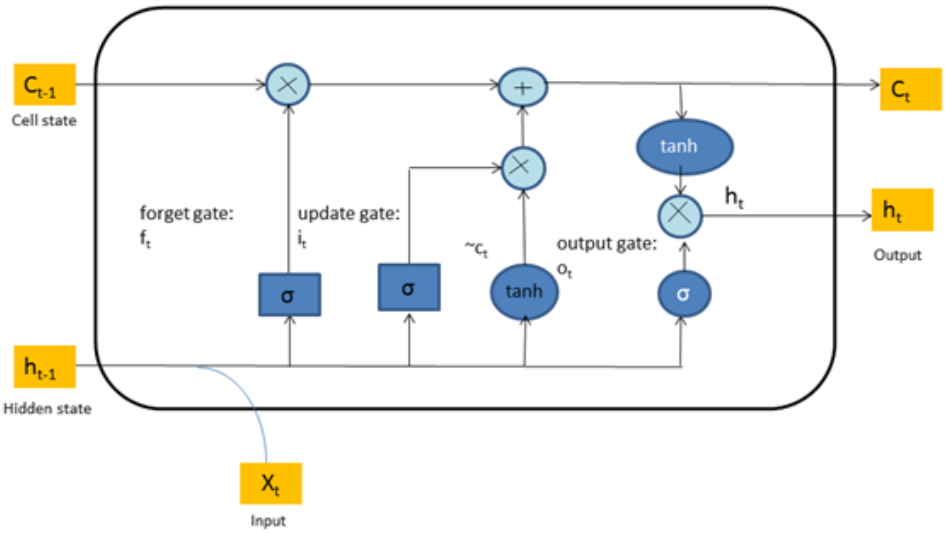


Figure 3.3: LSTM neural network structure

the backward states and vice-versa. Bi-LSTM uses two layers such that one layer performs the operations following the same direction of data sequence and the other layer applies its operation in the reverse direction of the data sequence.

3.4 Ensemble methods

Ensemble methods help to build a learning algorithm by combining base estimators to get optimal predictive model. The ensemble method tend to have higher accuracies, as more than one model is trained using a particular technique to improve the performance of the model and to reduce the overall error rate. Generally, it is used to enhance the predictability as well as to improve the robustness of a model. The ensemble approach is used because it is capable of boosting weak learners [87]. The significant improvement in the prediction by the use of an ensemble learning approach encourages the researchers to solve the problems of different fields. There are many more benefits of ensemble approach which include effective prediction results, selection of relevant features, a combination of data, incremental learning, class imbalance handling, and correction of errors. The ensemble approach uses divide and conquer method in which a complex problem is divided into multiple chunks that are easy to analyze and solve [88]. This approach has the advantage that the ensemble model can adapt to any diversity in the data more correctly as compared to single model. It suggests that the ensemble approach is more efficient than the single model. The progress of the ensemble approach depends upon the diversity in the individual model corresponding to unclassified instances. Polikar stated that there are four ways to attain this diversity [89]. Firstly, train the individual model with different

data chunk. Secondly, use different training parameters. Thirdly, use different properties to train the model and finally, combine different types of models. According to Dietterich [90], there are three reasons which conclude that the ensemble model is efficient than the single model. The first is that the training dataset does not always facilitate the required information to select one correct hypothesis. The second is that the weak models are not properly trained. The third is that the hypothesis space being searched might not get the proper target function while an ensemble model can produce a good approximation.

3.4.1 Need of ensemble learning

The process of ensemble learning is an effective approach to achieve a highly accurate model by combining less accurate ones. There is no one best machine learning model to solve all the cases of a given problem. Various techniques are used to improve the performance of machine learning models which include efficient pre-processing of the data, collect a large number of features and perform feature selection task to get relevant ones, explore different machine learning models. If results are not desirable, then we can combine the less accurate models. In an ensemble model, more than one opinions are there for a single instance. Thus, if one model fails to predict the correct output, there is a chance that the other models predict it correctly. There are two errors in the trained model, including bias error and variance. Bias error quantifies that on an average, the predicted values differ from the targeted values. High bias represents that the model is underperforming, which means it has missed some essential trends. On the other hand, the variance is used to quantify that the prediction produced on the same observation differ from each other. High variance means the model is overfitted and will produce adverse predictions on instances except for training dataset.

To deal with these errors, an ensemble approach is an efficient way. There are two methods to combine the models which include bagging and boosting. The ensemble model is a meta-algorithm, which is a combination of different models and in order to minimize the variance bagging ensemble approach is used, and in order to minimize the bias boosting ensemble approach is used.

3.4.2 Techniques of ensemble learning

Ensemble approach is beneficial to enhance the models performance. There are two ways to combine the different models and are explained below:

i. **Bagging**: Bagging means bootstrap aggregation, which is a simple and successful method to ensemble the models. It is used to improve the unstable classification

problems. For instance, weak models like decision tree can fluctuate when any training point changes its position and may become a different tree. This ensemble method can be applied to other models as well. Bagging method is beneficial for the vast and high dimensional datasets. It is introduced by Leo Breiman [91] to the variance of the model. In bagging, outputs of n models are aggregated which are generated by using N bootstrap sets. These sets are generated by using complete dataset via feature selection and random method with replacement. The parallel training of each model is possible because the training of each model is independent. In the end, voting or averaging of outputs is performed where each bootstrap set produces outputs.

ii. **Boosting** Boosting is introduced by Schapire [92], which is an ensemble technique to boost the performance of weak models and then group into a robust model. It facilitates the sequential training of the models. The first model is trained on the complete dataset while other models get trained by using training sets. These sets are based upon the output of the previous ones. The incorrect instances are extracted to increase their weights. Therefore, these instances have a high chance of appearing in the training dataset, which is used by the next model. By using this approach, different models are well trained on different sets of the data, which helps the ensemble model to produce enhanced results [93].

3.5 Dimensionality Reduction

In theory, the information provided by additional features should help to improve the models accuracy, however, in reality, these additional features increase the risk of overfitting, i.e., memorizing noise in the data rather than its underlying structure. For a given sample size, there is a maximum number of features above which the classifiers performance degrades rather than improves. This problem is called the curse of dimensionality, and the techniques for reducing of high-dimensional data intuitively into low-dimensional data fall into dimension reduction. One of the main aspects of the curse of dimensionality is the large size of the dataset. In fact, processing high-dimensional data is already a tough task in current scientific research. Feature selection is a technique where we try to map the high-dimensional data space into lower dimensional space with minor loss of information.

Table 3.2: Confusion matrix

	Predicted Healthy	*Predicted PD*
Actual Healthy	True Negative (TN)	False Positive (FP)
Actual PD	False Negative (FN)	True Positive (TP)

3.6 Performance Evaluation Measures

The performance of the ML methods applied in this thesis is evaluated from the confusion matrix as shown in Table 3.2. The parameters computed from the classification matrix include classification accuracy (calculated by eq. 3.17), precision (calculated by eq. 3.18), recall (calculated by eq. 3.19), F-score (calculated by eq. 3.20) and Receiver operating characteristic (calculated by eq. 3.21).

Accuracy Accuracy represents the overall success of positive and negative cases i.e. to distinguish between PD and non-PD cases and mathematically it can be represented in percentage. If the data is biased, the accuracy gets affected.

$$Accuracy(\%) = \frac{TP + TN}{TP + TN + FP + FN} * 100 \tag{3.17}$$

Precision Precision (P) or positive predictive value is an accuracy measure that classifies the events, which is equal to true positive/total number of predicted positives of a classifier.

$$Precision = \frac{TP}{TP + FP} * 100 \tag{3.18}$$

Recall Recall (R) or true positive rate is the potential of a classifier to accurately classify a patient who has disease. This is an accuracy measure that classifies the events, which is equal to true positive/total positive of a classifier.

$$Recall = \frac{TP}{TP + FN} * 100 \tag{3.19}$$

F-Score F-Score is the harmonic mean of both precision and recall.

$$F - Score = \frac{2 * Recall * Precision}{Recall + Precision} \tag{3.20}$$

Receiver Operating Characteristic Receiver operating characteristic curve (ROC curve) is a graphical method used to examine the overall performance of several methods.

It is a curve to plot the relationship between specificity and sensitivity. The area under the ROC (AUC) curve is a good summary of the overall performance of a specific classifier. AUC takes values between 0.5 and 1, which is an effective measure of the performance of a classifier and can be calculated as per Eq.3.21.

$$ROC = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{3.21}$$

3.7 Representation methods

Extracting features for both disease and unknown genes is one of the most significant tasks in identifying disease genes. In this research, we use protein sequences to characterize genes and used three representation methods to extract information encoded in proteins, such as Normalized Moreau–Broto Autocorrelation (NA) [24], Geary Autocorrelation (GA) [25] and Moran Autocorrelation (MA) [26]. These methods represent the neighboring impact between amino acids with a specific number of amino acids separated in their sequence using their specific physicochemical property. Also, make it possible to find patterns that run through whole sequence. These autocorrelations can be defined below as in Eq. 3.22 - 3.25.

Moreau-Broto autocorrelation for protein sequence are defined as:

$$AC(l) = \Sigma_{i=1}^{N-1} P_i P_{i+1}. \qquad l = 1, 2, 3, ...nlag \tag{3.22}$$

Normalized moreau-broto autocorrelation (NA) can be defined as:

$$NAC(l) = \frac{AC(l)}{N-1}. \qquad l = 1, 2, 3, ...nlag \tag{3.23}$$

Geary autocorrelation (GA) can be defined as:

$$GA(l) = \frac{\frac{1}{2(N-1)}\Sigma_{i=1}^{N-1}(P_i - \rho_{i+l})^2}{\frac{1}{N-1}\Sigma_{i=1}^{N}(P_i - \tilde{P}')^2} \qquad l = 1, 2, 3, ...nlag \tag{3.24}$$

Moran autocorrelation (MA) can be defined as:

$$MA(l) = \frac{\frac{1}{N-1}\Sigma_{i=1}^{N-1}(P_i - \tilde{P}')(P_{i+1} - \tilde{P}')}{\frac{1}{N-1}\Sigma_{i=1}^{N}(P_i - \tilde{P}')^2} \qquad l = 1, 2, 3, ...nlag \tag{3.25}$$

where l is lag of auto-correlation, P_i and P_{i+l} are the properties of amino acids, nlag is value of lag, $\tilde{P}'$ is considered property along sequence i.e. $\tilde{P}' = \frac{\Sigma_{i=1}^{N} P_i}{N}$

These representation methods used their physicochemical properties of amino acids to explain the neighboring effect between amino acids with other amino acids within a sequence. These methods help to gain relevant information, which is unknown in protein sequences. Since all the representation methods are established on physicochemical properties, hence we used twelve physicochemical properties in this paper to acquire further knowledge regarding amino acid sequences. We have utilized twelve physicochemical properties to provide more knowledge about amino acid sequence. The physicochemical properties used are polarity [27], residue accessible surface area in tripeptide [28], hydrophilicity [29], polarizability [30], solvation free energy [31], entropy of formation [32], partition coefficient [33], amino acid composition (AAC) [34], hydrophobicity [35], transfer-free energy [36], CC in Regression analysis [37], and graph shape index [38] respectively.

3.8 Statistical Tests

Statistical Tests are used to execute all statistical analysis during this research. We employ various statistical tests and analyses to evaluate the findings of the results obtained in this research.

3.8.1 The Kolmogorov-Sminrov test

Kolmogorov-Sminrov test for normality is selected to find the p-values indicating a normal ($p > 0.05$) or non-normal ($p < 0.05$) distribution for performance evaluation measures results [94]. It is a very efficient way to determine if two samples are significantly different from each other. The findings showed a mixture of both normal and non-normal distributed data so non-parametric statistical analysis can be used for testing between the resampling conditions because they have the advantage of making no assumption about data distribution.

3.8.2 Wilcoxon signed-rank test

A non parametric hypothesis test [95] used for testing between two groups for resampling conditions within over-sampling and under-sampling experiments. This test is used to determine whether the two independent samples from populations having the same distributions with same median without assuming normal distribution. As the Wilcoxon

signed-rank test does not assume normality in the data, it can be used when this assumption has been violated and the use of the dependent t-test is inappropriate. The calculation is based on order of observations in sample.

H_o: Distribution of two samples is equal.

H_1: Distribution of two samples is different.

Statistic: $U = R - \frac{n1(n1+1)}{2}$ [96], R is sum of ranks and n is number of observations in sample.

3.8.3 Friedman test

A non-parametric test for testing [97] between more than two groups for combi-sampling. This test is used to determine differences in samples. The Friedman test is the non-parametric alternative to the one-way ANOVA with repeated measures. It is used to test for differences between groups when the dependent variable being measured is ordinal. It can also be used for continuous data that has violated the assumptions necessary to run the one-way ANOVA with repeated measures.

H_o: Distribution of two samples is equal.

H_1: Distribution of two samples is different.

Statistic: $U = \frac{12}{ns(s+1)}\Sigma_{j=1}^{s} Rj^2 - 3n(s+1)$ [98], s = number of samples, Rj is sum of ranks for jth group.

3.9 Summary

In this chapter, the preliminaries and the methods used in this research have been discussed. It covers a brief introduction of the various machine learning methods, deep neural network, ensemble methods, and performance evaluation measures. Discussion of different representation methods and physicochemical properties of amino acids followed by statistical tests has also been done.

Chapter 4

Machine Learning ensemble method for the diagnosis of Parkinson's disease genes

Classification of the genes that cause or initiate different genes leading to diseases with neurological disorders like Parkinson's disease (PD), is a grave challenge in biomedical research. Detecting neurological disorders has a significant contribution to genetics, which require the deployment of machine learning methods that are still in their infancy. For exploring protein sequences (genes), computational analysis is vital since a manual comparison of multiple sequences results in impracticality. It helps to find a gene in the sequence and combine protein sequences into a class of similar sequences. This chapter compares multiple classification methods to identify Parkinson's disease using hydrophobicity and Amino Acid Composition as feature extraction methods. Classification methods are then combined to propose a 2-level ensemble method based on the false prediction rate.

4.1 Motivation

This section describes the motivation behind the diagnosis of Parkinson's disease genes, followed by the contributions. There are many methods that prioritized disease genes, either using Protein-Protein Interaction (PPI) data [9, 17] or functional similar data, such as gene expression profiles [20] and gene ontology [21]. The primary issue with these methods is how to find a suitable threshold for separating disease genes from unknown genes. Therefore, a sequence-translated method based on physicochemical properties of amino acids is used to construct a feature vector for each protein. In recent years, numerous approaches have been proposed to predict genes associated disease with sequence data [99, 100, 43, 53]. However, only a few of these are applied for PD gene identification. Yousef et al. [19] and Yang et al.[10] used their proposed method only to 50 Parkinson's disease related genes and achieved good performance. Since their database is very small and limited so their values are on higher side and better suited. It has also been analyzed that research work done in the field of gene identification using protein sequences is mainly restricted to Support Vector Machine (SVM). Though, a complete analysis of all

the Parkinson's disease genes has not yet been performed. In this chapter, several models are trained with a powerful machine learning approach on the basis of physicochemical properties of amino acids such as Hydrophobicity and Amino Acid Composition (AAC) to classify PD patients. Machine learning approaches, such as Naive Bayes (NB), Support Vector Machine (SVM), K-nearest neighbor (KNN), Neural Network (nnet), Logistic Regression, Decision tree (DT), Random Forest (RF) and ensemble methods are used to effectively diagnose Parkinson's disease. Several Machine Learning (ML) methods were extensively used for identification of disease genes, but such methods can't get results of an ideal classifier. Therefore, we proposed a 2-level ensemble method with majority voting based on the false positive rate to develop a classifier with significant improvements compared to existing methods. Following are the main contributions of this paper:

1. Collection of datasets with Parkinson's disease genes and non-Parkinson's disease genes has been performed.

2. Hydrophobicity and Amino Acid Composition (AAC) properties are evaluated to construct feature vectors.

3. Compare the performance of ten machine learning methods using correlation coefficient feature extraction technique to retrieve the best model based on performance metrics.

4. Proposed a 2-level ensemble method to identify genes that are responsible for Parkinson's disease. Comparative study is done to show the effectiveness of our proposed model.

4.2 Problem Statement

To identify Parkinson's disease genes, an efficient method is proposed that uses protein sequence dataset from Uniprot and NCBI databases. Hydrophobicity and AAC are the two physicochemical properties of amino acids that have been used to extract feature sets. Feature extraction method has been proposed to choose a more relevant feature set for analysis. A 2-level voting ensemble method is designed to build an efficient method. Figure 4.1 shows the architecture of the proposed method. The problem statement can be defined as follows: To identify PD genes from sequence S of proteins where S = s1, s2,.., sn and si represents the amino acids in a sequence. The job is to evaluate and analyze the best machine learning method to measure high efficiency in our proposed method.

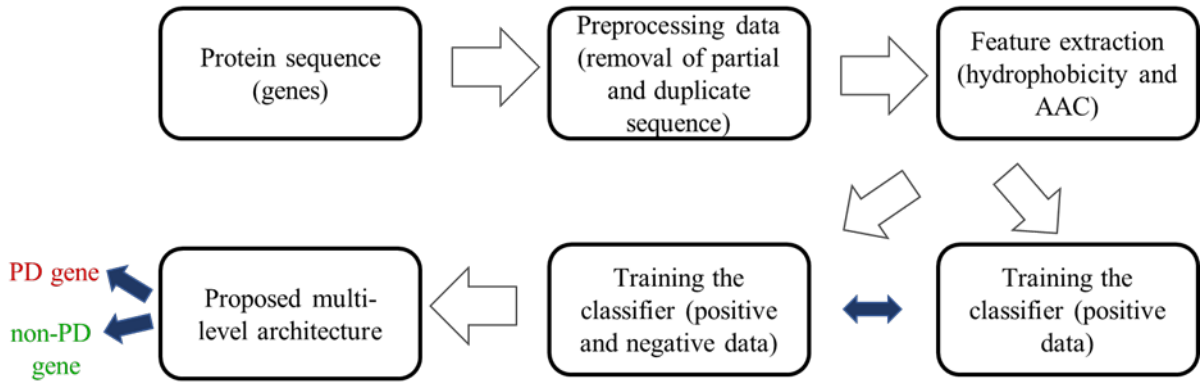


Figure 4.1: Architecture of the proposed ensemble method

4.3 Methodology

The steps involved in methodology are explained as shown in Figure 4.2. In the first step, protein sequences are collected from NCBI and UNIPROT database. The elimination of duplicates and partial protein sequences from the dataset are carried out in the cleansing phase. After cleansing, the sequences are transformed to numerical values using hydrophobic and AAC methods. Calculation of hydrophobicity score and AAC score using 38 hydrophobic scales and 20 amino acids to extract feature values for each protein sequence has been carried out in the extraction phase. In the classification phase, the data set is used to train machine learning models, with their tuning parameters. Fifth step consists of a confusion matrix that contains classification values for each protein sequence. The matrix generated in the fifth step is used to calculate performance parameters including accuracy, precision, recall, F-Score and ROC. In the comparison phase, a comparative score is generated by comparing each algorithm and then an ensemble method is proposed by merging these classifiers on the basis of majority voting. To measure the durability of the best predictive model, K-fold cross validation is used.

4.3.1 Data Collection

The dataset is formulated by obtaining human Parkinson's (PD) and non-Parkinson's (nPD) protein sequences from NCBI database and UNIPROT database. The extracted sequences were saved as a fasta file. The obtained dataset was then cleaned by eliminating duplicate and partial protein sequences. Total numbers of sequences are 1650, in which 640 are PD sequences and 1010 are nPD sequences. To extract features, 38 hydrophobic and 20 amino acid composition methods were applied on the 1650 sequences saved as fasta file format in which 640 are PD sequences and 1010 are nPD sequences. This

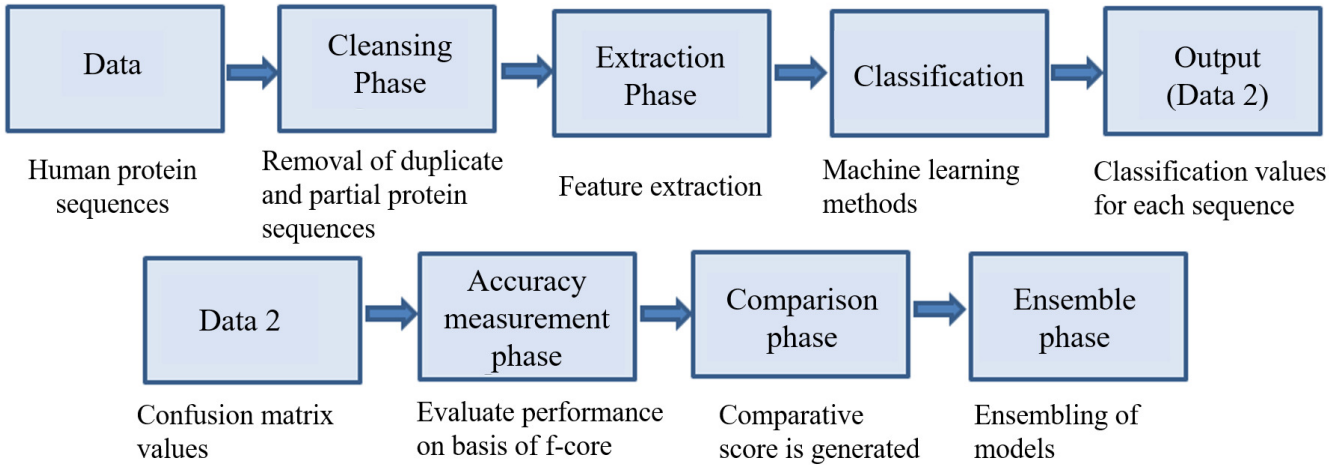


Figure 4.2: Steps involved in methodology

task was performed using built-in hydrophobic method libraries in R. After which an excel file was generated that indicated numerical values of the hydrophobicity and amino acid composition of all sequences corresponding to methods. A sample of the dataset of PD and non-PD genes having information such as accession number, gene, hydrophobic values (H1, H2....H38), AAC values (A1...A20) and Class is shown in Table 4.1.

4.3.2 Feature Extraction

Hydrophobicity is one of the important physicochemical properties of amino acids. Hydrophobicity is the property that is ostensibly repelled from a mass of water and hydrophilicity is the property of molecules that are attracted to water. Protein consists of

Table 4.1: Sample dataset of PD and non-PD genes having information such as accession number, gene, hydrophobic values (H1, H2....H38), AAC values (A1...A20) and Class

Accession No.	Gene	H1	H2	H38	A1	A20	Class
P00451	FA8	4.6393	0.2599	1.2595	0.0467	0.0531	Non-PD
P10636	TAU	4.5605	0.226	1.2987	0.0606	0.0926	PD
P35498	SCN1A	5.1207	0.4567	1.3796	0.0557	0.0686	Non-PD
P03886	NU1M	4.6213	0.2531	1.164	0.0668	0.0668	PD
Q99497	PARK7	4.7854	0.3301	1.2656	0.0907	0.0825	PD
P84022	SMAD3	6.427	0.9735	1.4383	0.0459	0.1781	Non-PD
P01130	LDLR	4.7134	0.2996	1.2865	0.0688	0.0728	Non-PD
P10721	KIT	4.0616	0.0422	1.1406	0.0789	0.047	Non-PD
P06213	INSR	4.3369	0.1341	1.2265	0.0626	0.0441	Non-PD
Q8IUQ4	SIAH1	4.6958	0.318	1.2973	0.0634	0.0655	PD
Q709C8	VP13C	5.3257	0.4885	1.3782	0.0428	0.0714	PD

both hydrophilic and hydrophobic amino acids. There are values that are used to outline the relative hydrophobicity of amino acid residues known as hydrophobicity scale [101]. This scale lies within a negative to positive range. Values that are in positive range means more hydrophobic are the amino acids located in that region of protein and values in negative range defined as not very hydrophobic. The hydrophobic methods are used to develop datasets for disease classification against machine learning models. The computed features include 38 hydrophobic and 20 amino acid composition methods as given in Table 4.2. This table describes the features along with frequency of each method in the dataset. These values are used for classification to train the machine learning models. Hydrophobicity of a sequence of amino acids can be calculated using Eq 4.1.

$$H = \sum_{1}^{i} P_d(p_i) \tag{4.1}$$

Where i is length of sequence of amino acid, pi is particular residue in amino acid sequence, Pd is hydrophobicity of pi and H is sum of all residues in a sequence.

Amino Acid Composition (AAC) is the fraction of amino acids of each type normalized with total number of residues. [102] used AAC to propose a computational technique for predicting protein to be monomeric or hetero-oligomers. AAC can be calculated using Eq 4.2.

$$AAC = \frac{\sum a_i * 100}{N} \tag{4.2}$$

Where i is residues of 20 amino acids, ai is the number of residues of each type, and N are the total number of residues.

4.3.3 Feature Selection

After the feature extraction phase, feature selection is done using correlation. The feature selection phase is carried out to extract relevant and best features from high-dimensional features (58 features). Since the variables are continuous in nature, Pearson correlation coefficients are used to create a matrix describing the dependencies in the data. Features with larger correlation coefficient values (r greater than 0.75) are first eliminated to remove the irrelevant information. Then the retrieved (25 features) are fed to the classification algorithm to improve the overall performance of methods. Table 4.2 shows the frequencies of each hydrophobic and amino acid feature.

Table 4.2: Feature Selection from Hydrophobic and Amino acid composition methods

Hydrophobic method			Amino acid composition	
Features	**Methods**	**Frequency**	**Features**	**Frequency**
H1	Aboderin	0.65	A1	0.19
H2	AbrahamLeo	0.803	A2	0.371
H3	Argos	0.725	A3	0.696
H4	BlackMould	0.802	A4	0.788
H5	BullBreese	0.429	A5	0.23
H6	Casari	0.805	A6	0.799
H7	Chothia	0.812	A7	0.71
H8	Cid	0.512	A8	0.723
H9	Cowan3.4	0.121	A9	0.365
H10	Cowan7.5	0.009	A10	0.496
H11	Eisenberg	0.755	A11	0.258
H12	Engelman	0.325	A12	0.128
H13	Fasman	0.753	A13	0.235
H14	Fauchere	0.127	A14	0.705
H15	Goldsack	0.006	A15	0.755
H16	Guy	0.523	A16	0.8
H17	HoppWoods	0.691	A17	0.412
H18	Janin	0.794	A18	0.805
H19	Jones	0.721	A19	0.639
H20	Juretic	0.214	A20	0.251
H21	Kidera	0.258		
H22	Kuhn	0.8		
H23	KyteDoolitte	0.632		
H24	Levitt	0.588		
H25	Manavalan	0.114		
H26	Miyazawa	0.723		
H27	Parker	0.102		
H28	Ponnuswamy	0.756		
H29	Prabhakaran	0.632		
H30	Rao	0.244		
H31	Rose	0.102		
H32	Roseman	0.951		
H33	Sweet	0.235		
H34	Tanford	0.803		
H35	Welling	0.621		
H36	Wilson	0.788		
H37	Wolfenden	0.129		
H38	Zimmerman	0.244		

4.4 Resampling Techniques

After the feature extraction, resampling techniques such as oversampling, undersampling and combi-Sampling are used in this work to deal with class imbalanced problem. Resampling is used to resample the smaller classes samples up to the majority class. Undersampling is used to resample the larger classes samples down to minority class. Combi-sampling is used to select a convergence point then utilize both over sampling and under sampling techniques based on relevant classes. This study focuses on the comparison of all machine learning methods with resampling techniques. For over-sampling, ADASYN and SMOTE are two different algorithms to be used within each technique. For under-sampling, Cluster Centroids (CC) and Near Miss (NM) are used within each technique. For combi-sampling, each algorithm from over and under sampling techniques are used in combination with each other to produce four approaches.

4.5 Proposed ensemble method

The proposed multi-voting ensemble method consists of feature representation, feature extraction, classifiers and voting phases. The features represented from hydrophobicity and AAC are then fed to extract relevant features using correlation coefficient. After that, ten different machine learning classifiers are used to classify the PD associated genes. The main motivation behind the ensemble method is to analyse the independence between the base learners. Two levels are performed to propose an ensemble method. At 1st level, sets of three machine learning methods are combined on the basis of their false positive rate to generate a new voting classifier. From 1st level, three voting classifiers are retrieved from each of three methods respectively. At the 2nd level, false predictions from all these three voting classifiers are merged to get the final model. The block diagram of the proposed method is shown in Figure 4.3. Ensemble method involves combining multiple model predictions to give better performance than an individual model. The main goal of this work is to enhance the model's performance by rechecking its false prediction rate. Based on the minimum false positive (FP) rate, the ML models are combined to overcome the weakness of the existing individual models. The minimum the FP ratio, the more accurate the model will be able to classify the genes as PD. Based on the FP ratio, the models are selected and merged to give an appropriate prediction. First block at level 1, ADA, RF AND GBM methods are merged together based on their false positive rate to build the Voting Classifier (V1). Second block at level 1, Voting Classifier (V2) is built after merging CART, SVM and neural net models on the basis of their false positive rate. Third block at level 3, pcannet, Nb and Logistic methods are merged together on

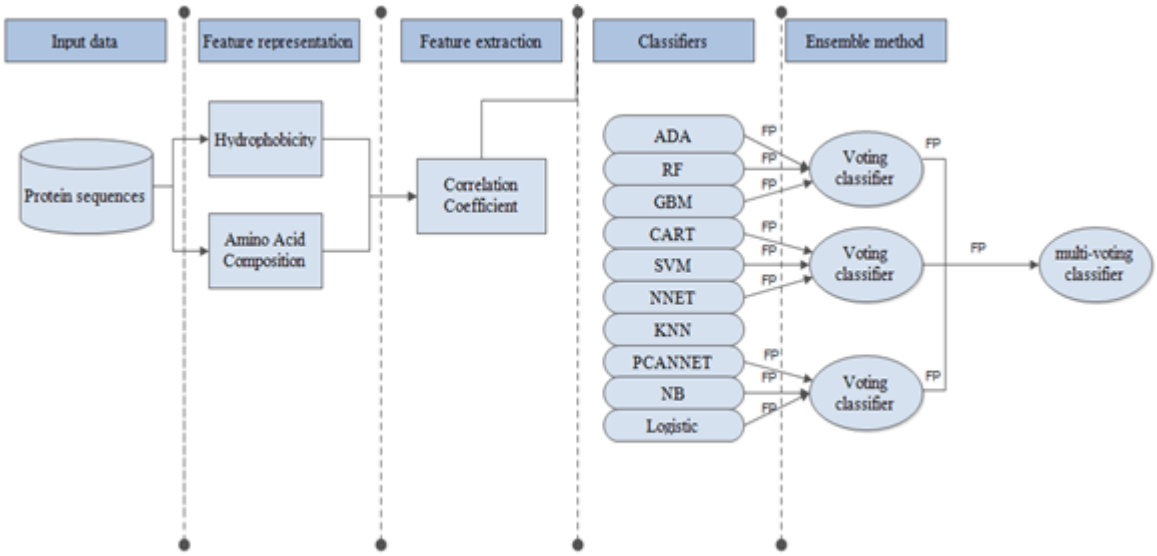


Figure 4.3: Block diagram of proposed multi-voting system

the basis of their false positive rate to build the Voting Classifier (V3). The machine learning methods are merged on the basis of the minimum false positive rate to overcome the weakness of the existing individual models. At level 2, a new voting classifier is built based on the false positive rates of proposed three classifiers (V1, V2 and V3). The values of false positive rate has been given in Table 4.4. The new multi-voting classifier is built to give final prediction of proposed method.

Table 4.3: Performance Evaluation of Machine Learning Models

SNo	Model Name	Accuracy	Precision	Recall	F-Score	ROC
1	ADA	86.69	84.76	86.34	85.54	89.69
2	RF	88.39	87.97	86.34	87.14	90.34
3	GBM	86.69	85.19	85.71	85.44	87.35
4	NB	88.95	87.2	88.82	88	90.32
5	Neural net	85.84	85.35	83.23	84.27	90.01
6	Pcannet	88.67	88.54	86.34	87.42	90.65
7	Decicion tree	86.4	84.66	85.71	85.36	88.41
8	SVM	88.39	87.97	86.34	87.14	90.34
9	Logistic	85.84	86.27	81.99	84	88.25
10	Knn	87.1	88.39	85.09	86.7	89.39

4.6 Performance Evaluation

The proposed ensemble and individual methods have been implemented in RStudio, which is open source software licensed under GNU GPL. Processor used is Intel core i7 processor with 16 GB RAM. Keras library of Python is used to implement Deep Neural Network.

Table 4.4: False Positive Rate for each machine learning model

SNo	Model Name	False Positive Rate
1	ADA	0.065
2	RF	0.081
3	GBM	0.064
4	NB	0.097
5	Neural net	0.085
6	Pcannet	0.091
7	Decicion tree	0.084
8	SVM	0.089
9	Logistic	0.120
10	Knn	0.178

The five performance measures are used to classify PD and non-PD patients. Various machine learning algorithms used for analysis include Adaboost, Random Forest, Gradient Boosting, Naïve Bayes, Classification and Regression Trees, neural network, Support Vector Machine, k nearest neighbour and Logistic Regression. The performance of these models is analysed using 5-fold cross validation. A comparative analysis is shown in Table 4.3. The boxplot distribution of best four models with 5-fold cross validation is shown in Figure 4.4 and the comparison of the average accuracy for all models is presented in Figure 4.5.

4.7 Result Analysis

Classification Accuracy, Precision, Recall, F-measure and area under ROC are evaluated for each model on 5-fold cross validation. Table 4.4 shows the comparison of ten classification methods and the accuracy of the best four models are illustrated in box-plots as shown in Figure 4.4. On average, RF, SVM, NB and PCANNET achieved the highest average accuracy rate (88.6%), followed by KNN (87.1%), ADA, GBM and DT (86.6%), Neural Net and LR (85.6%) as shown in Figure 4.5. The notches in the box plots are overlapped in all cases, so we can say that true medians do not differ. In terms of data dispersion, the interquartile ranges (upper and lower quartiles) have also gained a higher value for box plot. Since the average performance of all methods is good, there are no outliers in the plot. However, for all methods, the skewness pattern is not straightforward. All the boxes shown in Figure 4.5 seem skewed, because the median line markers face the top of the box. Therefore, box indicates that accuracy of these methods has higher upper values than lower. These box plots also achieve the highest value of upper quartile among all the mentioned methods. Among 10 classifiers, the RF, SVM, NB and PCANNET are found to be performed better than other methods. Neural networks also

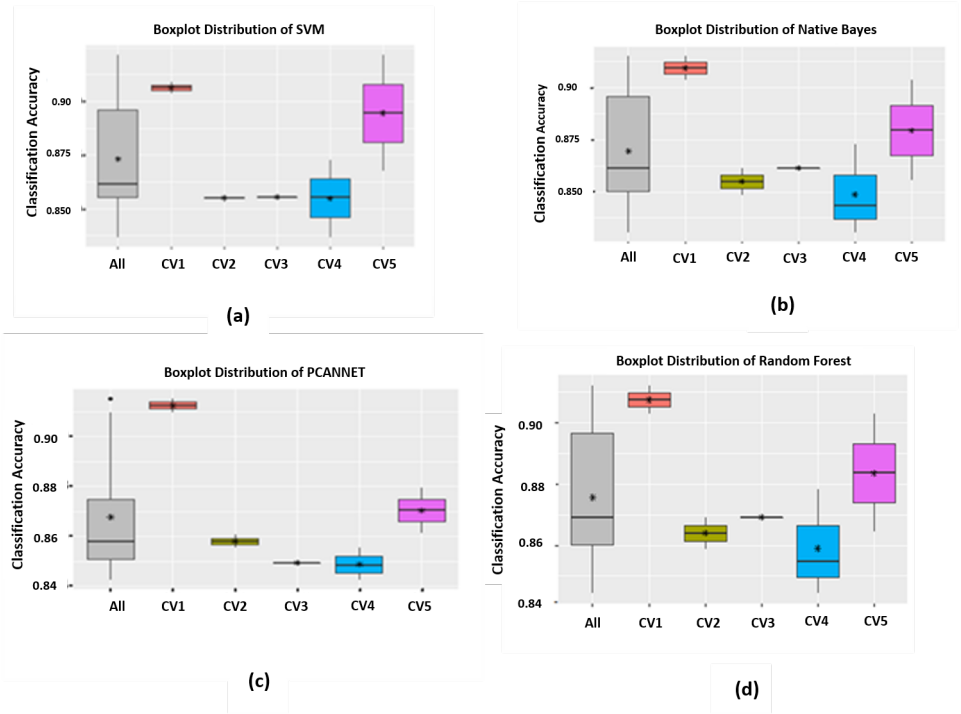


Figure 4.4: Boxplot distribution for accuracy with 5-fold Cross Validation (CV) a SVM, b Naïve Bayes, c PcaNNet and d Random Forest. CV1 is Cross Validation1, CV2 is Cross Validation2, CV3 is Cross Validation3, CV4 is Cross Validation4, and CV5 is Cross Validation.

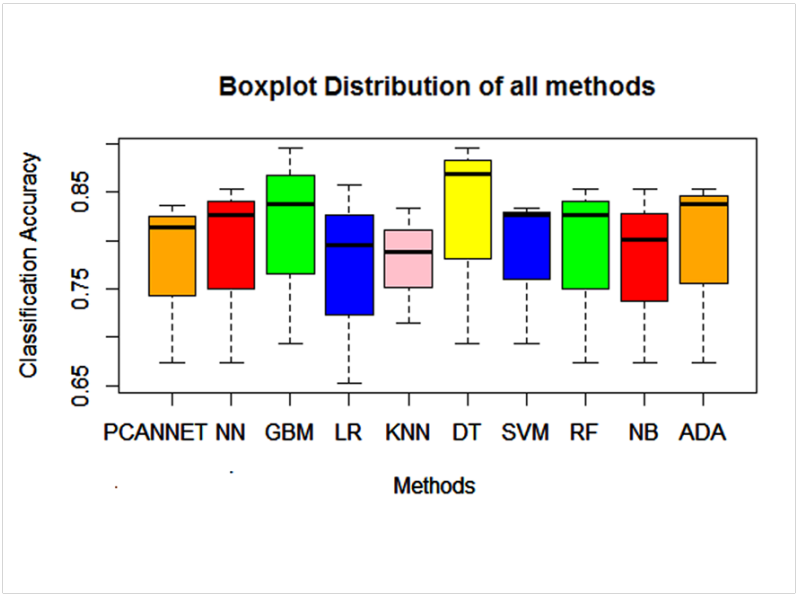


Figure 4.5: Boxplot distribution for classification accuracy by ten methods with average results from 5-fold CV

not showing good performance, probably due to small amounts of data.

Table 4.5: Statistical analysis results from wilcoson signed-rank test for resampling methods.

Resampling	**Optimal Strategy**	**AUC**	**F-Score**
Over Sampling	p(resampling)	5.45E-02	4.88E-02
	p(control)	0.00E+00	1.00E-03
Under Sampling	p(resampling)	0.00E+00	0.00E+00
	p(control)	0.00E+00	0.00E+00
Combi Sampling	p(resampling)	0.00E+00	0.00E+00
	p(control)	0.00E+00	2.70E-02

Table 4.6: Statistical analysis results from friedman test for combi-sampling, and Wilcoson signed-rank test for under and over sampling.

Combi-Sampling	**AUC**	**F-Score**
ADASYN-CC	3.00E-03	0.00E+00
ADASYN-NM	0.00E+00	0.00E+00
SMOTE-CC	1.00E-01	2.70E-02
SMOTE-NM	0.00E+00	1.80E-02
Over-Sampling	**AUC**	**F-Score**
ADASYN	0.00E+00	4.88E-02
SMOTE	5.45E-02	1.00E-03
Under-Sampling	**AUC**	**F-Score**
CC	0.00E+00	0.00E+00
NM	0.00E+00	0.00E+00

4.8 Statistical Analysis Results for Resampling Techniques

The statistical methods used here is Friedman and Wilcoxon signed-rank test to find the stochastic nature of algorithms used in the resampling techniques. Over sampling and under sampling results are determined from direct significance testing between conditions within each resampling experiment. SMOTE and ADASYN for oversampling and, Combi-sampling utilized combination of over-sampling and under-sampling methods. The statistically significant differences between conditions are shown with Area Under Curve (AUC) and F-Score, defined as $p < \alpha$, where $\alpha = 0.05$ for under sampling, over sampling. Table 4.5 shows the results for optimal strategy within re-sampling experiments and Table 4.6 shows the results of Statistical analysis from wilcoson signed-rank test and friedman test for resampling methods. Under-sampling shows inferior performance for both AUC and F-Score measures. Both AUC and F-Score values are highest in case of

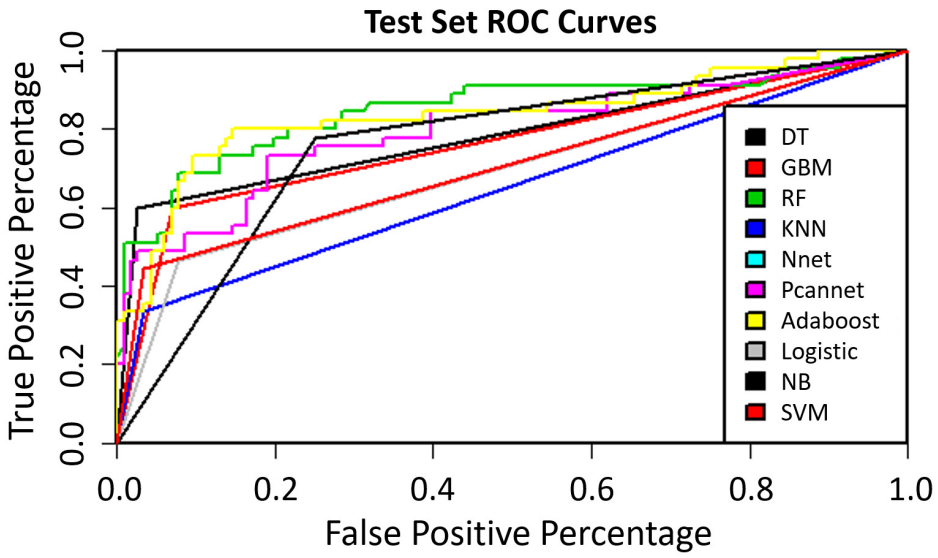


Figure 4.6: TPR vs. FPR for all classification methods

over sampling. In over-sampling, AFDASYN is the best-performing resampling condition by rank from ROC-AUC and SMOTE from F-Score metric. However, in both parameters, neither resampling condition had significant differences to each other. Table 4.5 shows that the p-value for AUC and F-Score is lower than the alpha (0.05). As a result, the null hypothesis is rejected, and we will accept the alternative hypothesis that refers to there is a difference in the performance within the various resampling methods on same dataset. The combi-sampling would not be the suitable technique if there is small gap between minority-majority class and if dataset is small.

4.9 Proposed ensemble method evaluation

In order to evaluate the performance of a proposed ensemble method on an imbalanced dataset, the choice of evaluation parameters plays an important role. As the same class has less effect on accuracy as compared to the majority class, therefore, accuracy is not a suitable parameter to calculate performance. Therefore, other metrics should be used to access the classifier performance on imbalanced datasets. F-measure, Precision and Recall are the most relevant criteria to evaluate imbalanced data. F-Score is taken into consideration to average out the results of both precision and recall. The precision metric helps to identify genes that are predicted diseased genes and belongs to the PD genes category also. The recall metric helps to identify genes that are predicted diseased genes but actually belong to both healthy and diseased genes category. F-Score considers both false positive and false negative. Accuracy is considered as a good parameter when both

false negative and false positive values are closer to each other. Comparative analysis of all machine learning methods has been shown in Table 4.3. The purpose of ROC is to find the increase in false positive rates (FPR) with an increase in true positive rates (TPR) at various threshold settings. The performance of methods at various thresholds is shown in Figure 4.6. The evaluation parameters of the proposed ensemble method are depicted in Figure 4.7

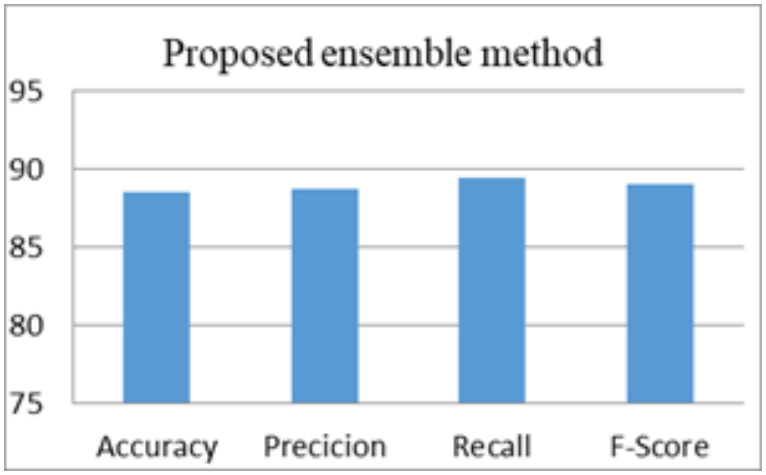


Figure 4.7: Evaluation parameters of proposed ensemble method

4.10 Comparison of proposed method with Deep Neural network (DNN) based methods

In this section, we have compared our proposed ensemble method with Deep Neural Network (DNN) based methods to identify PD genes. Various results have been discussed in this section to validate our proposed ensemble method. The performance analysis of all the deep learning methods (MLP, LSTM and Bi-directional LSTM) is evaluated in terms of accuracy, precision, recall, F-Score and area under curve (AUC) as shown in Table 4.7 and Figure 4.9. As we have to predict the percentage of actual positive instances that means we have to find genes that are responsible for Parkinson's disease, so we can analyze our results on the basis of Recall values. The proposed ensemble method achieves a recall value of 89.50% which outperforms MLP, LSTM and Bi-directinal LSTM with 5.21%, 7.69% and 4.38% respectively. Figure 4.8 shows the training accuracy versus testing accuracy of the different state-of-the-art deep neural networks. From the accuracy plot, we can see that the both training and testing data accuracy goes on increasing with epochs but in case of bidirectional, data is overfitted. From the loss plot, we can see that losses are reduced drastically for MLP. For LSTM and Bidirectional LSTM, loss values varies for both training and testing datasets. We can see that performance at train time

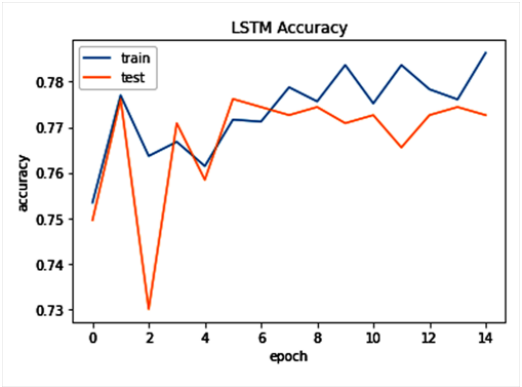
Table 4.7: Comparative analysis of proposed ensemble method with deep network-based methods

Methods	**Hidden layers**	**Accuracy**	**Precision**	**Recall**	**F-Score**	**AUC**
MLP	4	76.55	74.17	84.29	78.9	83.23
LSTM	4	77.26	74.35	81.81	77.9	82.54
Bi-directional LSTM	3	78.86	77.33	85.12	81.03	83.08
Proposed ensemble method	-	88.5	88.81	89.5	89.15	85.57

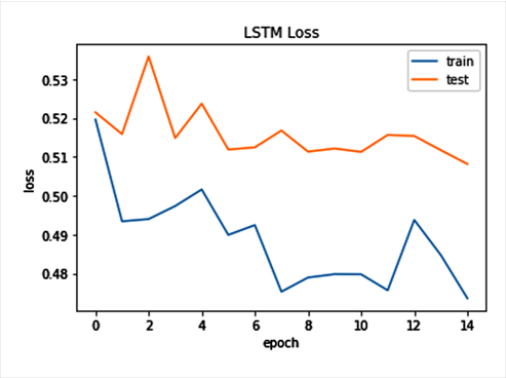
continually decreases while validation performance rises suggesting overfitting. From the plots, we can conclude that LSTM does not perform well on our dataset. The proposed ensemble method also achieves highest AUC value as depicted in Table 4.7 and Figure 4.9. Hence, the proposed ensemble method can be used further for predicting Parkinson's disease genes.

4.11 Comparison of proposed method with state-of-the-art

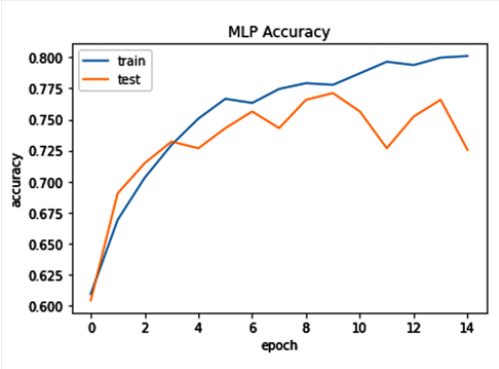
We have compared our proposed ensemble method with state-of-art methods, including, Xu's method [1], Smalter's method [34], ProDiGe method [3], PUDI method [4] and EPU method [35]. The major difference between the previous methods and proposed ensemble method is about the genes knowledge that is used to extract features. Most of the previous methods used topological properties of protein-protein interaction data, functional data, gene ontology and gene expression data to extract features. All these methods may not be implemented properly because they rely on prior knowledge, which may become expensive, or information about certain training and testing genes may not be available or may contain noisy information. In this work, we have extracted features from protein sequences which are universally available for all genes. For example, Xu's method considered only topological features but ignored function related features to identify disease genes. In the PUDI method, some features are not available for some genes. In this work, protein is defined by a sequence which is available for all proteins. Table 4.8 shows the comparison between proposed ensemble method and five state-of-art methods. It can be observed from the table that the proposed ensemble method shows higher F-Score value than existing methods. Also, the proposed ensemble method gives significant improvement for all three parameters, including, Precision, Recall and F-Score. From the results, we can conclude that fusing multiple classifiers is more accurate and potent than individual classifiers.



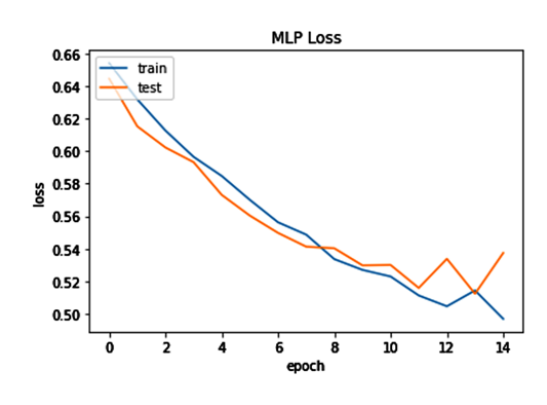
(a)



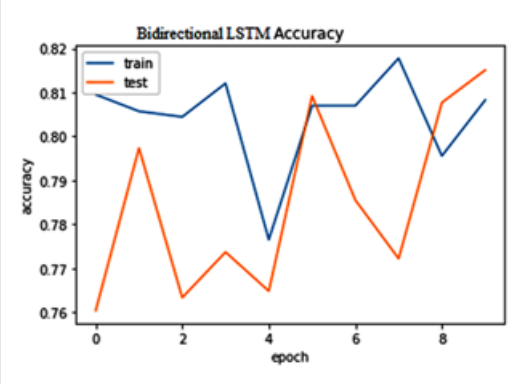
(b)



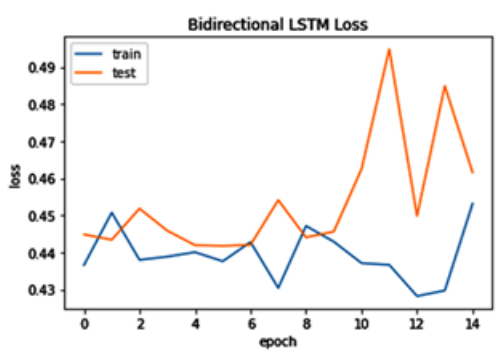
(c)



(d)



(e)



(f)

Figure 4.8: Training accuracy versus testing accuracy for (a) LSTM, (c) MLP, (e) Bidirectional LSTM, and Training loss versus testing loss for (b) LSTM, (d) MLP, (f) Bidirectional LSTM; at different epochs.

4.12 Summary

In this chapter, various machine learning classifiers were analysed to find the conditions under which a particular model were outperforms others for identifying Parkinson's dis-

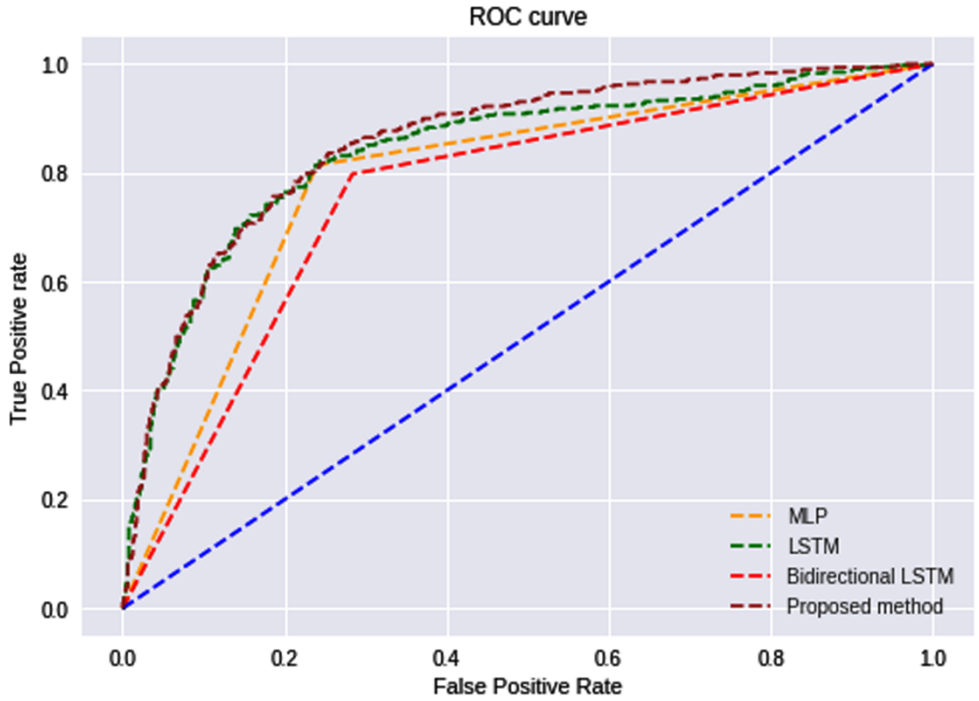


Figure 4.9: ROC curve of deep learning methods and proposed ensemble method methods

Table 4.8: Comparison of existing methods with proposed ensemble method for neurological disorders

Method	**Classifier**	**Precision**	**Recall**	**F-Score**
Xu's method [1]	KNN	59.7	66.7	63
Smalter's method [34]	SVM	60.6	65.9	63.1
ProDiGe [3]	SVM	63.1	74	68.1
PUDI [4]	Multi-level weighted SVM	70.3	80.1	74.9
EPU [35]	KNN, Naïve Bayes, SVM	78.2	80.4	78.6
Proposed ensemble method	Ensemble	88.8	89.5	89.1

ease genes. The protein sequences were transformed into numerical feature vectors using hydrophobic and AAC properties of amino acids. Feature extraction method has been proposed to choose a more relevant feature set for analysis. Machine learning approaches, such as Naive Bayes (NB), Support Vector Machine (SVM), K-nearest neighbor (KNN), Neural Network (nnet), Logistic Regression, Decision tree (DT), Random Forest (RF) and ensemble methods were analyzed on the basis of performance measures to effectively diagnose PD. After analyzing the classifiers performance measures, the focus was to utilize the strengths of one model to complement the weakness of another. So, an ideal multi-level

voting model was proposed, which integrates various ML models based on their FP rates to retrieve a new voting classifier to retrieve better prediction analysis. The developed model helps to solve the trade-off issue between accuracy and efficiency. We found that ensemble method could better model the classification problem for disease gene prediction as it achieved significantly better results than the state-of-the-art methods. Given that many machine learning problems in biomedical research do involve positive and unlabeled data instead of negative data, we believe that the performance of machine learning methods for these problems can possibly be further improved by adopting a Positive Unlabeled approach [32], [103] as we have done here for disease gene identification. For future work, we will consider to integrate more biological resources [104], such as gene expression data etc. In addition, we may explore more complicated machine learning methods to better model the positive and unlabeled data distributions.

Chapter 5

Parkinson's disease genes identification using N-semble method

5.1 Introduction

Healthcare is significantly impacted by the discovery of the link between human genetic disorders and the genes that cause them. A growing number of genes have been identified as disease-causing genes as a result of the quick advancement of biomedical research. Based on their genetic correlations to those known disease-causing genes, machine learning techniques can be used to find new disease-causing genes. In particular, positive unlabeled learning approaches have recently been put forth to develop a classification model where known genes are treated as positive training set P and unknown genes are treated as unlabeled set U (instead of negative set N) because unknown genes contain unidentified disease genes. There are many experimental approaches that have been introduced in recent years to identify disease genes from a vast number of candidate genes. These techniques differ in the genomic data type used to generate feature vectors, such as protein-protein interactions (PPI) [9],[17], gene expression profiles [20], and protein and biological functions. Unfortunately, all these techniques are based on the information of proteins achieved from protein domains, gene ontology, and, PPI data. Hence, might have not been implemented accurately as information is incomplete, noisy, and time-consuming. Data that can be used for all proteins and has a significant role in solving issues such as protein-protein interactions [105], [9], and predicting subcellular locations [106] is the protein sequences. An identification method is proposed in this work using protein sequences to identify proteins (genes) that are responsible for having PD. Earlier, for machine learning-based studies, the task of disease gene prediction was usually approached as a binary classification problem, where training data consisted of positive and negative training samples. This is a limitation of the binary classification-based methods because the negative training samples should be actual non-disease genes. However, it is nearly impossible for biomedical researchers to construct this set. Thus, more practical approaches have been proposed. To overcome this limitation, recent studies proposed to learn the scoring function only from known and unknown gene set, by

formulating the problem as positive (P) and unlabeled (U) learning (PU learning) [31]. In this chapter, we designed a novel neural network-based ensemble (n-semble) method to integrate multiple PU learning classifiers for more accurate and robust disease gene prediction. The Artificial Neural Network (ANN) is trained in a unique way to ensemble the multiple model predictions. In comparison to both individual PU learning classifiers and state-of-the-art approaches, the proposed method, in our observation, has produced noticeably better outcomes. We are able to significantly reduce the possible bias and risk of individual predictions by merging the outputs of numerous PU learning classifiers. As a result, the expected errors using our ensemble approach can be expected to be substantially decreased. Protein sequences and a wide range of computational classifiers can be effectively predicted by the suggested strategy. As more trustworthy biological data sources and potent computational classifiers become available in the future, we anticipate that our method can be substantially enhanced by incorporating these new, high-quality biological data sources and computational techniques.

5.2 Motivation

An n-semble model is proposed in this work to build an efficient classifier with significant enhancement over existing methods for identifying disease genes. Several machine learning models have been used for identification of Parkinson's disease genes. The major contributions are as follows:

1. Collection and statistical analysis of Parkinson's and non-Parkinson protein sequences (genes) from NCBI, Ensembl and Uniprot databases were performed.

2. Twelve physicochemical properties of amino acids were applied to generate features with Geary Autocorrelation, Normalized moreau-broto autocorrelation and moran autocorrelation representation methods.

3. The t-Distributed Stochastic Neighbor Embedding (t-SNE) feature reduction method was used to extract relevant features from high-dimensional feature vectors.

4. Six machine learning methods were evaluated for gene identification to find the best model based on performance measures and a neural network-based ensemble model was put forth.

5. The performance of the proposed n-semble method was analysed using parameters like precision, recall and F-Score and the comparative study was conducted to show the effectiveness of the proposed model.

5.3 Problem Statement

To address the issue of binary classification as there is no negative data available, an appropriate Positive Unlabeled learning technique is applied to extract negative samples from unknown samples. A novel N-semble model has been proposed to develop an efficient disease genes identification system. Mathematically, a problem statement can be represented as follows: To identify gene G = PD, nPD for a protein sequence S where S= $a_1, a_2, ..., a_n$ and a_i represents the amino acid in a sequence. The task is to build a appropriate classification method using machine learning methods on positive unlabeled data to compute high efficiency in our proposed system.

5.4 Proposed method

To identify Parkinson's disease genes, sequence representation methods with physicochemical properties of amino acids are chosen to improve the efficiency of existing machine learning classifiers. In this paper, we have employed twelve physicochemical properties of amino acids to represent the amino acid features. Therefore, relying on more physicochemical properties will allow us to discover more information about the interactions. However, the increased characteristics lead to generate more features for each protein, which is why we have normalized the output feature vector, instead of concatenating the feature vector of two proteins. A novel n-semble method is proposed to develop an efficient disease gene identification method.

The proposed n-semble method for identifying PD genes has been described in this section. The proposed approach consists of four steps: (1) Adopting twelve physicochemical properties to transform corresponding protein sequences into feature vectors; (2) t-Distributed Stochastic Neighbor embedding (t-SNE) is applied to reduce dimensionality (3) differentiating negative samples from unknown genes; (4) modelling features using n-semble method. The proposed method architecture is depicted in Figure 5.1.

5.4.1 Extracting features from protein sequences

Extracting features for both disease and unknown genes constitutes one of the most significant tasks in identifying disease genes. This work applies protein sequences to characterize genes and used three representation methods to extract information encoded in proteins, such as Normalized Moreau–Broto Autocorrelation (NA) [107], Moran Autocorrelation (MA) [108] and Geary Autocorrelation (GA) [109]. These methods represent adjacent influences between amino acids that have a specific ratio of amino acids apart

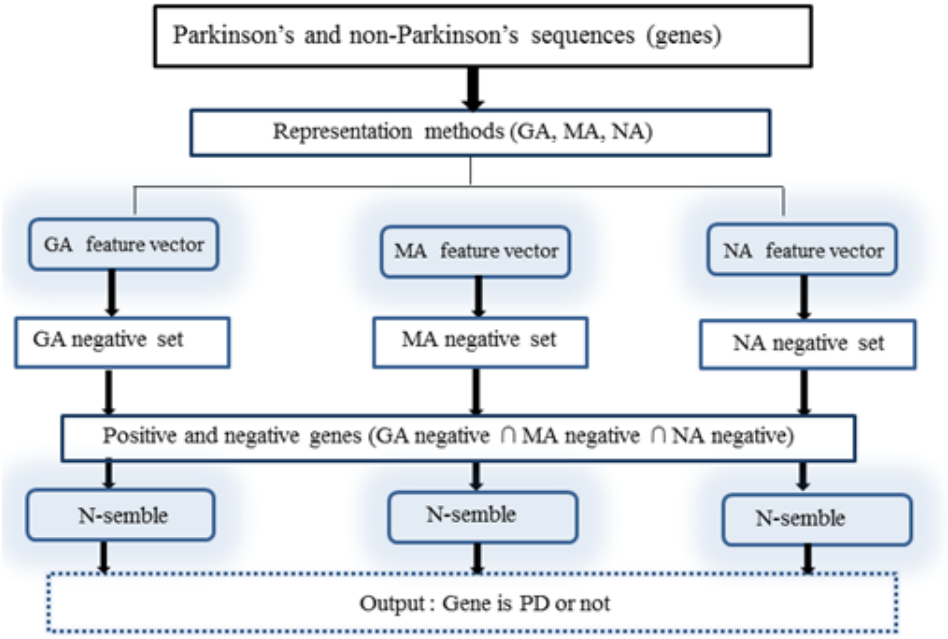


Figure 5.1: Architecture of proposed method

in the sequence using their particular physicochemical property. Similarly, it is possible to find patterns throughout the sequence. We have used these representation methods to avoid missing significant information regarding the protein sequences. Moreover, the selected methods are being used in several other works [39] also and have an advantage over other methods. We used twelve physicochemical properties of amino acid to attain more information regarding the amino acid sequence. The physicochemical properties include polarity (pol) [110], residue-accessible surface area (RAS) in tripeptide [111], hydrophilicity (hy phil) [112], polarizability (polz) [113], solvation-free energy (SFE) [114], entropy of formation (entrp) [115], partition coefficient (PC) [116], amino acid composition (AAC) [110], hydrophobicity (hy phob) [117], transfer-free energy (TFE) [118], Correlation Coefficient (CC) in Regression analysis [119], and graph shape index (GSI) [120]. Further, the min-max normalization method is considered to normalize the original values of the physicochemical properties. The normalized values of these physicochemical properties for each of amino acid (A-Y) are shown in Table 5.1.

5.4.2 Effectiveness of physicochemical properties

Since representation methods are based on the physicochemical properties, so our objective is to understand which physicochemical property has more impact in disease gene

identification. In this regard, twelve different GA feature vectors have been built. The first one is generated using all physicochemical properties, and the five other ones are generated by deleting one of the physicochemical properties, respectively. Figure 5.2 illustrates the effectiveness of physicochemical properties in disease gene identification. From the results, we could observe that the deletion of the hy-phil property reduces the F-Score dramatically. While, deleting the others has a small negative effect on the performance. However, the deletion of any physicochemical properties will decrease the performance.

Table 5.1: Normalized values of physicochemical properties

	POL	RAS	HY-PHIL	POL-ZAB	HY-PHOB	SFE	AAC	CC	GSI	TFE	PC	EOF
A	0.4939	6492	0.6009	0.3118	0.5491	0.3236	0.564	0.4697	0.599	0.4121	0.4009	0.2933
C	0.4498	0.5717	0.4832	0.4401	0.5024	0.3218	0.5284	0.3475	0.562	0.3236	0.5087	0.2901
D	0.3728	0.6047	0.4179	0.2146	0.4728	0.158	0.5186	0.3965	0.5159	0.3995	0.3693	0.2514
E	0.4425	0.635	0.2676	0.3707	0.5331	0.3168	0.5892	0.4205	0.526	0.4682	0.4376	0.3083
F	0.396	0.6876	0.4544	0.4539	0.4181	0.388	0.4401	0.3815	0.5786	0.2829	0.3541	0.2655
G	0.5671	0.7023	0.512	0.4337	0.7802	0.2965	0.7777	0.4015	0.4695	0.5605	0.5172	0.3611
H	0.3364	0.5164	0.5201	0.4416	0.4001	0.3185	0.4208	0.3528	0.6075	0.2744	0.22	0.3118
I	0.4286	0.6246	0.6031	0.4445	0.4597	0.2866	0.4661	0.3501	0.547	0.2916	0.4461	0.2579
K	0.3155	0.4872	0.5784	0.493	0.3641	0.3038	0.3742	0.3285	0.4999	0.3519	0.4141	0.2694
L	0.2813	0.5541	0.503	0.5562	0.3162	0.3696	0.318	0.2812	0.4504	0.2919	0.4126	0.2711
M	0.4521	0.6842	0.5479	0.52	0.4633	0.5097	0.4918	0.2708	0.5198	0.4121	0.2806	0.2739
N	0.3942	0.6144	0.751	0.3679	0.4453	0.2707	0.4713	0.3482	0.4971	0.3672	0.3911	0.2732
P	0.3528	0.4531	0.5024	0.43	0.4077	0.2605	0.4627	0.3244	0.3955	0.333	0.3647	0.2675
Q	0.347	0.6151	0.5335	0.3759	0.3929	0.2866	0.4136	0.3226	0.5062	0.3617	0.2535	0.2487
R	0.3506	0.5536	0.4662	0.4221	0.4088	0.2647	0.4368	0.3318	0.5104	0.3212	0.3458	0.2634
S	0.4163	0.5525	0.501	0.2969	0.4258	0.2452	0.4561	0.3731	0.3117	0.2534	0.3248	0.2981
T	0.3936	0.5981	0.4245	0.1972	0.4304	0.2273	0.4353	0.3199	0.515	0.4013	0.4832	0.2881
V	0.447	0.6924	0.4565	0.2857	0.4619	0.2755	0.4881	0.335	0.5401	0.3413	0.3952	0.2977
W	0.2983	0.6543	0.433	0.3315	0.4322	0.3181	0.4399	0.364	0.565	0.4133	0.3421	0.2873
Y	0.4578	0.7251	0.4432	0.2991	0.4728	0.3355	0.5119	0.3448	0.458	0.3203	0.2834	0.2526

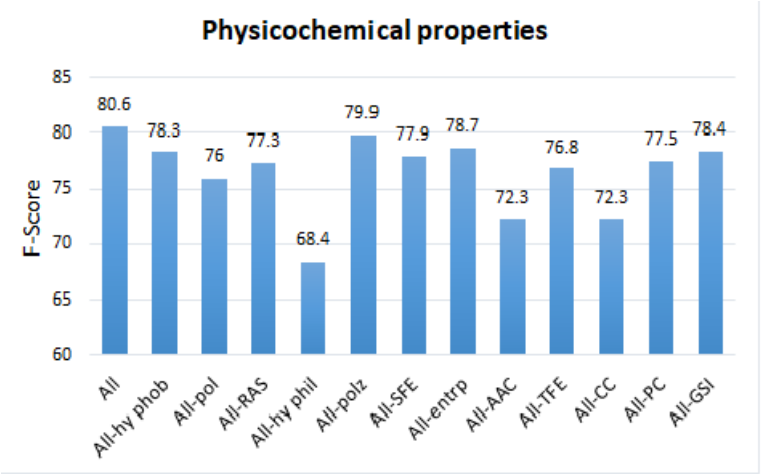


Figure 5.2: Importance of physicochemical properties

5.4.3 t-Distributed Stochastic Neighbor Embedding (t-SNE)

We applied the t-SNE dimensionality reduction method to find the most important and useful features from high-dimensional data. t-SNE is a non-linear dimensionality reduction approach that can identify observed clusters created on similarity of data points with multiple features, thereby detecting patterns in data. It is better suited for converting high-dimensional data into a space of low-dimensional data in such a way that similar instances are modelled by closed instances and dissimilar instances are modelled by distant instances. It helps to calculate the probability similarity of points in both high and low-dimensional space. Therefore, it is used to find similar features that retain most of the information and remove redundant information [121]. t-SNE minimizes the KL (KullbackLiebler) divergence between the two distributions with respect to location of instances in a map. The existing features in the dataset may have some irrelevant features from the high dimensional data (360 features), which may decrease the performance of classifiers and result in poor accuracy. Thus a proper feature extraction technique for pre-processing of input data is required. The t-SNE extracted features have less correlation and less redundancy among the features, which consequently increases the internal representation of a dataset. These modified data representations improve the performance of classifiers. Table 5.2 shows the number of t-SNE extracted features with sequence-represented methods.

5.4.4 Extracting negative samples

After extracting relevant features with the above feature reduction method, it becomes a requirement to develop a classifier for PD genes identification. For this, reliable negative

Table 5.2: Number of t-SNE extracted features for different representation methods

Method	**Number of features**	**t-SNE features**
Geary Autocorrelation (GA)	360	65
Moran Autocorrelation (MA)	360	60
Normalized Moreau–Broto Autocorrelation (NA)	360	71

genes need to be extracted from unknown genes to construct a method together with positive and reliable negative genes. We propose an Algorithm 5.1 for selecting negative genes from unknown samples (US). The algorithm comprises of six steps. First, initialize the negative set as an empty set. Second, compute the positive set (PS) of all positive proteins for each of MA, GA and NA representation methods, respectively. Third, compute the unknown set and assign any one value of the representation method. Fourth, compute the similarity between an unknown sample (US) and positive mean (Pm). The Jaccard similarity metric has been evaluated to calculate distance between each protein and positive mean. Fifth, find the reliable (r) negative genes from US by selecting the sample farthest from the positive mean vector for each feature vector. Finally, the resulted genes acquired by intersection of selected negative genes are considered as a reliable negative set. Table 5.3 shows the comparison of three different distance measures to compute the negative samples. Yang et al. [16] applied Euclidean distance to find the negative set from unknown samples. Euclidean distance gives better results only if positive data show identity covariance. Therefore, the distance measure directly affects the cogency of the extracted negative set. According to the results shown in Table 5.3, the Jaccard matric yields better results when compared with the other two methods.

Table 5.3: Comparison between distance metrics

Distance Methods	**Precision**	**Recall**	**F-measure**
Jaccard	88.9	90.9	89.8
Cosine	84.5	86.6	85.5
Euclidean	80.6	83.8	82.1

5.4.5 n-semble

The proposed n-semble model helps to improve the performance of the classifiers. The motivation behind the proposed method is to analyze the interdependence between base learners. Two levels are proposed to perform the experiment as discussed below.

- **Level 1:** Three machine learning models are selected based on F-Score to train the neural network.

Algorithm 5.1 Selection of negative genes from unknown samples

1. Procedure negative samples(S)
2. for each: $rn_i \varepsilon$ RN
   Set (rn_i, NULL);
   Until RN!= NULL;
3. for each: $ps_j \varepsilon$ PS — Δ PS is a positive set
   $V_{ps(j)}$ <-ASSIGN($V^{MA}{}_{PS}$,$V^{NA}{}_{PS}$, $V^{GA}{}_{PS}$); — Δ Assign any one value
   Total_PS = Total_PS + $V_{ps(j)}$ until j=n;
4. for each: $us_j \varepsilon$ US; — Δ US is a positive set
   $V_{us(j)}$ <-ASSIGN($V^{MA}{}_{US}$, $V^{NA}{}_{US}$, $V^{GA}{}_{US}$); — Δ Assign any one value
   until j=n
5. for each: $V_{ps(j)} \varepsilon$ PU
   $\text{dist}^{MA}{}_{PU} = \text{dist}(V^{MA}{}_{PU}, V^{MA}_{ps(j)})$
   $\text{dist}^{GA}{}_{PU} = \text{dist}(V^{GA}{}_{PU}, V^{GA}_{ps(j)})$
   $\text{dist}^{NA}{}_{PU} = \text{dist}(V^{NA}{}_{PU}, V^{NA}_{ps(j)})$
   until ps(j)! = NULL
6. NP_D^{MA} = Sort(dist$^{MA}{}_{PU}$); — Δ Negative protein set
7. NP_D^{GA} = Sort(dist$^{GA}{}_{PU}$)
8. NP_D^{NA} = Sort(dist$^{NA}{}_{PU}$)
9. Select NP_D set
   NP^{MA} = Select(NP_D^{MA}(1:r))
   NP^{GA} = Select(NP_D^{GA}(1:r))
   NP^{NA} = Select(NP_D^{NA}(1:r))
10. RN_Set = $NP^{MA} \cap NP^{GA} \cap NP^{NA}$
11. End Procedure

- **Level 2:** A neural network is trained using the prediction results of the top three selected models and actual values of these predictions.

The block diagram of the proposed n-semble model is depicted in Figure 5.3. It is comprised of 3 parts: 1) Data Partition, 2) Data Classification, training and testing of selected models, and 3) Training and testing of a neural network. Three feature representation methods (GA, MA and NA) are used to represent features from collected protein sequences. The features retrieved from the feature extraction phase are fed to various classification algorithms. The data obtained is split 75% into training and 25% into testing phase. In the second phase, various classification algorithms are applied such as Random Forest (RF), Support Vector Machine (SVM), Adaboost, Decision Tree (DT), Xgboost and Gradient Descent to identify genes. The models are selected based on their prediction performance. The top three models on the basis of F-Score are then integrated to form an ensemble method to achieve high efficiency. In the last phase, the predictions of the selected model are used as training data, and the actual predicted values are taken as target values. The predicted actual data set is applied to train the neural network, the

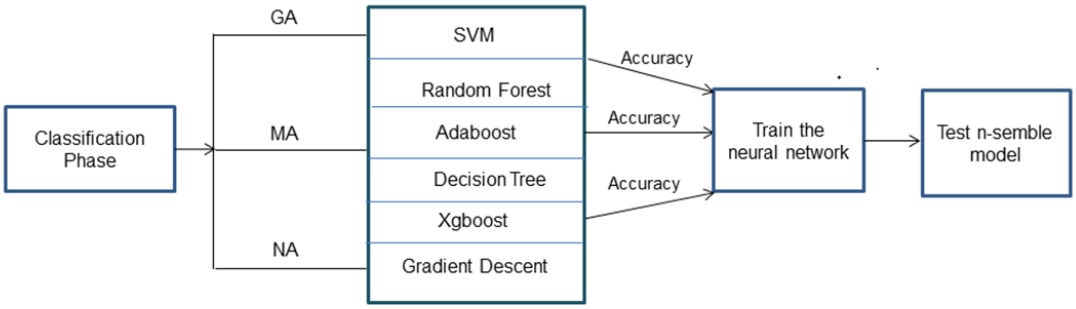


Figure 5.3: Block diagram of proposed n-semble method

size of which is 20% of the data set. Adaboost, Random Forest and Xgboost are selected as top three models based on highest F-Score among other methods. The training data create the relationship between the actual model and the predicted values of the top three models by calculating the weights required for the predictions assigned to each model. Each network has a hidden layer containing ten hidden units. The size of the input layer is the same as the number of attributes in training data and sigmoid activation function is adopted for the output layer.

5.5 Experimental Results

The performance of the proposed n-semble method on the imbalanced data set is evaluated in this section. First, we investigated the impact of three sequence representation methods such as GA, MA and NA on the performance of n-semble method. Additionally, an optimal number of features retrieved through the t-SNE method has been reviewed and optimised. Then, the effect of several machine learning methods has been evaluated, and on the predictions of top three models, a neural network is trained to develop an ensemble method. Finally, our method and another disease gene identification method were compared to confirm the method effectiveness. The ML methods applied in this work are Support Vector Machine (SVM), Random Forest (F), Adaboost, Decision Tree (DT), Xgboost, Neural network and Gradient Descent. Table 5.4 shows the values of various performance measures, i.e. Precision, Recall, and F-Score for comparative analysis of the ML models with whom experimented. The top three models, random forest, adaboost and xgboost were selected on the basis of the highest F-Score used to generate an ensemble method. The prediction values evaluated by means of each selected model are used as training data for the neural network, and the actual prediction values are used as target data. As shown in Table 5.4, the proposed ensemble method outperforms other methods with Precision (88.9%), Recall (90.9%) and F-Score (89.8%). It was ob-

served that the proposed method outperforms the Adaboost by 2.8%, Xgboost by 4.5% and Random Forest by 5.4%. To evaluate the predictive performance of all methods, the ROC (Receiver Operating Characteristic) curve is plotted. The performance of True Positive Rate (TPR) versus False Positive Rate (FPR) at various thresholds for the selected methods is shown in Figure 5.4. Random Forest outperforms other methods with the area of 83.6% under ROC.

Table 5.4: Comparative analysis of machine learning methods

Model	Precision	Recall	F-Score
SVM	80.1	81.2	80.6
Random forest	83.1	85.9	84.4
Gradient Descent	82.6	85.6	84
Xgboost	84.2	86.5	85.3
Adaboost	85.8	88.4	87
Decision Tree	80.2	80.7	80.4
N-semble	88.9	90.9	89.8

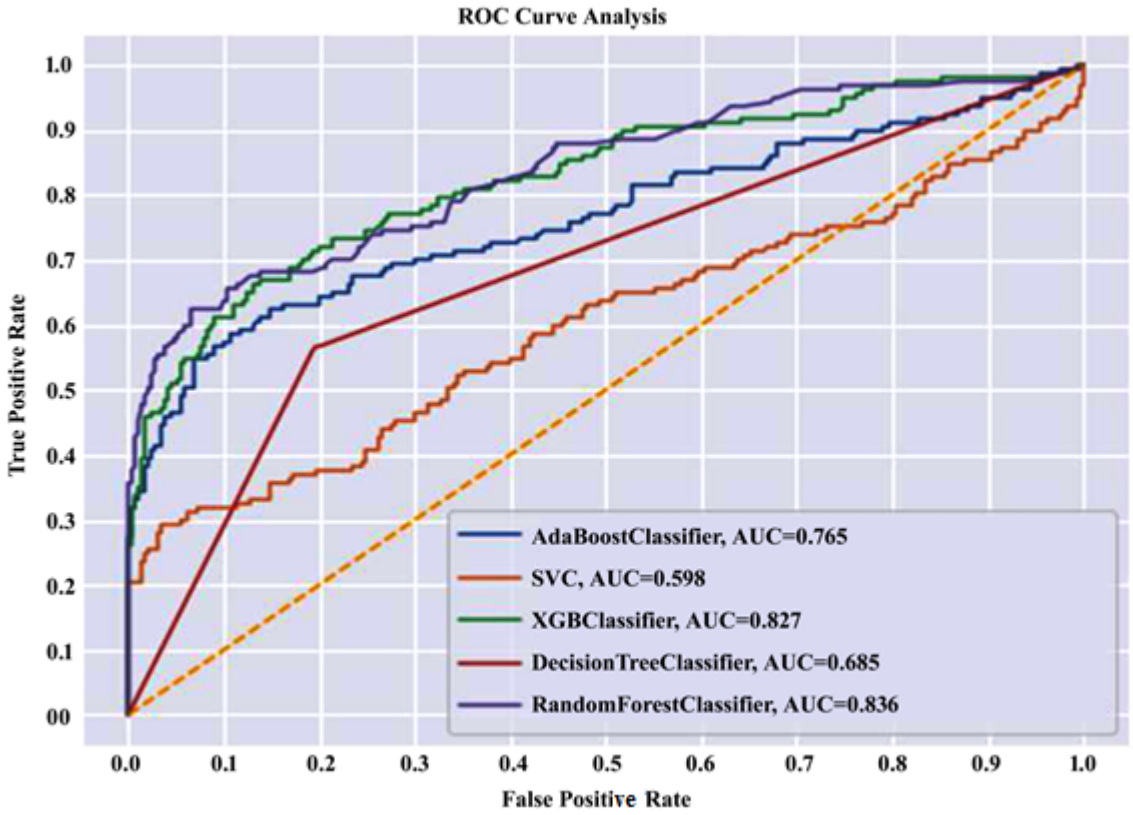


Figure 5.4: True Positive Rate versus False Positive Rate of selected methods

5.5.1 Validation of proposed method

We have performed k-fold cross validation due to its simplicity and randomness property to validate the robustness of the proposed ensemble model. The process cross-validated different samples of equal size, k times [122]. We have considered the k value as 5 i.e. the model is trained and tested 5 times. Use random data samples of the same size to train and test the model each time, and then compare the results. Figure 5.4 shows the k-fold values of the top three, selected models. Figure 5.4(a) shows the results of precision on selected and n-semble methods. From the graph, we can infer that the n-semble method curve lies on top of other models. It indicates that the proposed model is robust as the plot shows a straight line. This means that the accuracy of the model has nothing to do with the given data sample and remains unchanged for a fixed data set. Figure 5.4(b) shows the results of recall values on other selected and n-semble methods. Figure 5.5(c) demonstrates the results of F-Score values on other selected and n-semble methods. The curve of the n-semble method is above the selected methods, which proves that the gain in F-Score is robust, that is, independent of the samples obtained from the dataset.

5.5.2 Comparison with state-of-art techniques

In this section, the n-semble method is compared with five state-of-the-art methods, namely, Xu's method [38], Smalter's method [27], ProDiGe [32], PUDI method [31] and EPU method [10]. The comparison between the proposed method and other existing methods is shown in Table 4.8. The F-Score of proposed approach averages 5.4%, 6.4%, 10.1%, 16.9%, 22.8% and 23.4% higher than with Yousef's method, SFM, PUDI, ProDiGe, and Smalter's method, respectively, for imbalanced datasets. The key difference between these other methods and the proposed one is the previous information used to generate features. The protein sequences were realized as the important information to generate features in this paper, and in previous methods, prior information was affected by noise. The second issue centres on the extraction of negative samples from unknown genes. Smalter's method [27] considered unknown or candidate genes as negative samples, while ProDiGe [32] randomly used multiple negative samples of unknown genes. The PUDI [31] method applied the Euclidean metric to find distance between each gene features and a positive vector. However, the feature vector generated by PUDI consists of noisy data. Yousef's method [19] applied only positive data to train a model, which is an ineffective approach. In this paper, we find the Jaccard distance metric the most reliable method for selective negative genes from unknown samples.

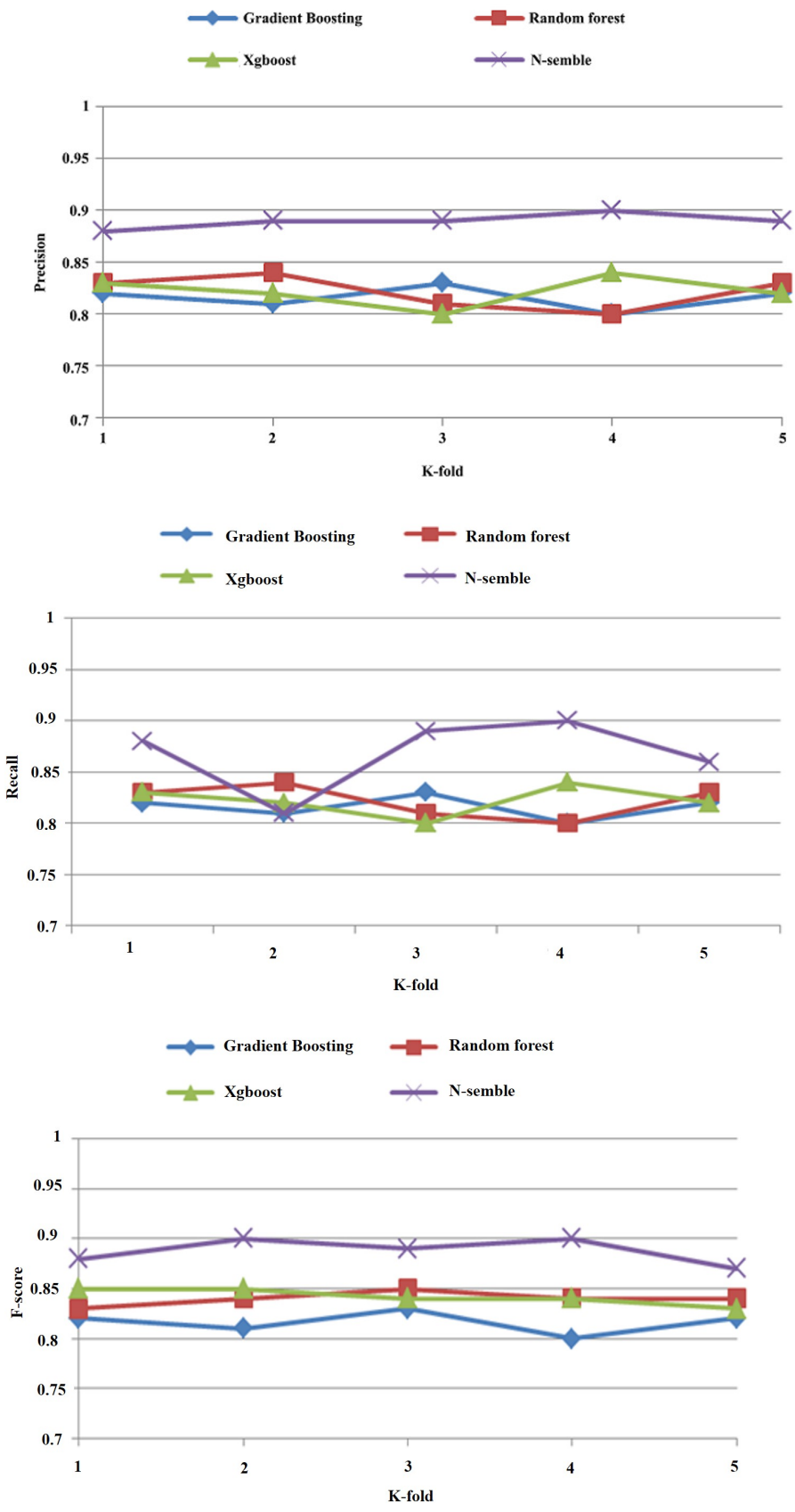


Figure 5.5: K-fold cross-validation for (a) Precision, (b) Recall, (c) F-Score

Table 5.5: Comparative analysis with state-of-art methods

Method	Precision	Recall	F-Score
PUDI	70.3	80.1	74.9
ProDiGe	63.1	74	68.1
Smalter's method	60.6	65.9	63.1
Xu's method	59.7	66.7	63
EPU	78.2	80.4	78.6
N-semble	88.9	90.9	89.8

Table 5.6: Comparative analysis with existing ensemble approaches

Method	Precision	Recall	F-Score
Wcomb	69.6	83.5	75.9
Ucomb	65.3	74.7	70
EPU	78.2	80.4	78.6
N-semble	88.9	90.9	89.8

5.5.3 Comparison with existing ensemble approaches

We contrast our proposed approach with three ensemble baselines: the first uses a weighted combination based on the component models' accuracy, while the second applies a uniform combination of the three models trained separately and the last is EPU method [10]. Table 5.6 evaluates four ensemble approaches. N-semble regularly outperforms other ensemble methods in a significant way, indicating that either uniform or weighted combination cannot balance component classifiers with the appropriate weights. Due to equal weighting of all components, uniform combination (UComb) has the weakest performance. The weighted combination (WComb), on the other hand, generally compares the single classifier situation to the ensemble classifier scenario.

5.6 Case Study: Predicting novel disease genes

Given a particular disease class, the set of confirmed disease genes are obtained from UNIPROT. Using all these disease genes as a positive training sets, we perform experiments by applying our proposed n-semble algorithm to predict novel disease genes from all the unlabeled gene set. As a case study, we have selected two significant disease categories: Parkinson's and cancer. In order to rank all the genes, we use the probability predicted by trained model. Based on the literature survey, we discover that several of these genes have been linked to Parkinson's disease. We first applied our method to discover novel disease genes for Parkinson's disease. We then search literature to check whether any of these predicted disease genes are really related to PD. We found that

nine predicted genes, namely, DHDDS, PARK2, PICK1, MT-ND4, NDUFB5, NDUFA6, CIC, TRIM63 and ATP5A1 have been reported to be associated with PD. Table 5.7 lists these candidate disease genes and related literature evidence to support their association with PD. We used our technique to find new disease genes for Parkinson's disease first. After that, we do a literature review to see if any of these predicted disease genes are actually connected to Parkinson's disease. Nine anticipated genes DHDDS, PARK2, PICK1, MT-ND4, NDUFB5, NDUFA6, CIC, TRIM63, and ATP5A1 were discovered to be linked to PD. These potential disease genes are listed in Table 5.7 along with any relevant research linking them to Parkinson's disease. Our proposed algorithm predicts 20 unlabeled genes as candidate disease genes for cancer illness gene prioritisation. PHF10, ABCB10, PRDX4, PRDX5, SIGLEC7, and SRPK1 are six of them that have been linked to cancer-related illnesses. These potential disease genes are listed in Table 5.8 along with any supporting research from the literature.

5.7 Summary

In this work, we proposed a n-semble method to identify genes associated with Parkinson's disease. To specify the conditions under which a classification method outperforms other classifiers is a key question in machine learning. This chapter, introduced various methods, including Support Vector Machine (SVM), Random Forest (RF), Adaboost, Decision Tree (DT), Xgboost and Gradient Descent for genes identification. After evaluating and analysing the classification methods, more emphasis is placed on exploiting the strengths of a model to complement the weaknesses of another. Therefore, an n-semble method was proposed which trained a neural network in a special way and integrated three classification methods based on their F-Score to ensemble the predictions and to achieve more accurate predictive analysis. On the basis of various performance measures, results from the proposed n-semble method show enhanced performance compared to state-of-the-art works. We have adopted protein sequences based on previous knowledge to extract features. GA, MA AND NA representation methods with twelve physicochemical properties of the amino acids are adopted to convert protein sequences into numerical feature vectors. Consequently, t-SNE is applied to extract relevant features. We found that physicochemical properties of amino acids would be highly beneficial in extracting features. Compared with the previous methods on unbalanced datasets, the proposed n-semble method improves the F Score. In this work, we have shown that the GA representation method is characterized by a higher success rate than other representation methods. Therefore, in the future, we will consider using a single GA feature vector to combine multiple different classifiers to improve the classification. We will also use this

Table 5.7: Predicted novel PD genes by n-semble method

Gene	**Reference**
DHDDS	Piccolo G., et al. (2021). Complex neurological phenotype associated with a de novo DHDDS mutation in a boy with intellectual disability, refractory epilepsy, and movement disorder. Journal of Pediatric Genetics, 10(03), 236-238.
PARK2	Kilarski L. L., et al. (2012). Systematic review and UK-based study of PARK2 (parkin), PINK1, PARK7 (DJ-1) and LRRK2 in early-onset Parkinson's disease. Movement Disorders, 27(12), 1522-1529.
PICK1	He J., et al. (2018). PICK1 inhibits the e3 ubiquitin ligase activity of parkin and reduces its neuronal protective effect. Focant, M. C. and Hermans, E. (2013). Protein interacting with C kinase and neurological disorders. Synapse, 67(8), 532-540.
MT-ND4	Ouzren N., et al. (2019). Mitochondrial DNA deletions discriminate affected from unaffected LRRK2 mutation carriers. Annals of Neurology, 86(2), 324.
NDUFB5	Talebi, R., et al. (2016). Parkinson's disease and lactoferrin: analysis of dependent protein networks. Gene Reports, 4, 177-183.
NDUFA6	Talebi, R., et al. (2016). Parkinson's disease and lactoferrin: analysis of dependent protein networks. Gene Reports, 4, 177-183.
CIC	Vurture G., et al. (2018). Outcomes of intradetrusor onabotulinum toxin A injection in patients with Parkinson's disease. Neurourology and Urodynamics, 37(8), 2669-2677. Peng J., et al. (2019). Predicting Parkinson's disease genes based on node2vec and autoencoder. Frontiers in genetics, 10, 226.
TRIM63	Ham S. J., et al. (2021). Loss of UCHL1 rescues the defects related to Parkinson's disease by suppressing glycolysis. Science Advances, 7(28), eabg4574.
ATP5A1	Shamir R., et al. (2017). Analysis of blood-based gene expression in idiopathic Parkinson disease. Neurology, 89(16), 1676-1683.

Table 5.8: Predicted novel cancer genes by n-semble method

Gene	**Reference**
ABCB10	Liang H. F., et al. (2017). Circular RNA circ-ABCB10 promotes breast cancer proliferation and progression through sponging miR-1271. American journal of cancer research, 7(7), 1566.
PRDX4	Park S. Y., et al. (2020). PRDX4 overexpression is associated with poor prognosis in gastric cancer. Oncology letters, 19(5), 3522-3530. Tiedemann K., et al. (2019). Exosomal release of L-plastin by breast cancer cells facilitates metastatic bone osteolysis. Translational oncology, 12(3), 462-474.
PRDX5	Bai F., et al. (2020). PCAT6 mediates cellular biological functions in gastrointestinal stromal tumor via upregulation of PRDX5 and activation of Wnt pathway. Molecular Carcinogenesis, 59(6), 661-669. Li S.et al. (2018). The prognostic values of the peroxiredoxins family in ovarian cancer. Bioscience reports, 38(5).
SIGLEC7	Rodriguez E., et al. (2021). Sialic acids in pancreatic cancer cells drive tumour-associated macrophage differentiation via the Siglec receptors Siglec-7 and Siglec-9. Nature communications, 12(1), 1-14.
SRPK1	Van Roosmalen W., et al. (2015). Tumor cell migration screen identifies SRPK1 as breast cancer metastasis determinant. The Journal of clinical investigation, 125(4), 1648-1664. Wang F., et al. Involvement of SRPK1 in cisplatin resistance related to long non-coding RNA UCA1 in human ovarian cancer cells. Neoplasma, 62(3), 432-438.
PHF10	Wet M., et al. (2010) Preparation of PHF10 antibody and analysis of PHF10 expression gastric cancer tissues. Journal of Xiao Bao Yu Fen Zi Mian Yi Xue 26(9): 874-6.137 Li C., et al. (2012) MicroRNA-409-3p regulates cell proliferation and apoptosis by targeting PHF10 in gastric cancer. Cancer Lett 320(2): 187-97
ABCB10	Tang, L., et al. (2009) Exclusion of ABCB8 and ABCB10 as cancer candidate genes in acute myeloid leukemiaLetter to the Editor. Leukemia 23: 1000-2
PRDX4	Lee S.U. et al. (2008) Involvement of peroxiredoxin IV in the 16alpha-hydroxyestrone-induced proliferation of human MCF-7 breast cancer cells. Cell Biol Int 32(4): 401-5 Park H.J. et al. (2008) Proteomic profiling of endothelial cells in human lung cancer. J Proteome Res 7(3):1138-50

Gene	Reference
PRDX5	Enqman L. et al. (2003) Thioredoxin reductase and cancer cell growth inhibition by organotellurium compounds that could be selectively incorporated into tumor cells. Bioorg Med Chem 11(23): 5091-100. McNaughton M., et al. (2004) Cyclodextrin-derived diorganyl tellurides as glutathione peroxidase mimics and inhibitors of thioredoxin reductase and cancer cell growth. J Med Chem 47(1): 233-9. Enqman L., et al. (2000) Water-soluble organotellurium compounds inhibit thioredoxin reductase and the growth of human cancer cells. Anticancer Drug Des. 15(5): 323-30.
SUGLEC7	Ito A. et al. (2001) Binding specificity of siglec7 to disialogangliosides of renal cell carcinoma: possible role of disialogangliosides in tumor progression. FEBS Lett.
SRPK1	Hayes, G.M., et al. (2007) Serine-arginine protein kinase 1 overexpression is associated with tumorigenic imbalance in mitogen-activated protein kinase pathways in breast, colonic, and pancreatic carcinomas. Cancer Res. 67(5): 2972-80.

method in the prediction of other neurological diseases.

Chapter 6

LSTM and MLP based multi-feature extraction for diagnosis of Parkinson's disease genes

Machine learning methods can be applied to discover new disease genes based on the known ones. Existing machine learning methods typically use the known disease genes as the positive training set P and the unknown genes as the negative training set N (non-disease gene set does not exist) to build classifiers to identify new disease genes from the unknown genes. However, such kind of classifiers is actually built from a noisy negative set N as there can be unknown disease genes in N itself. As a result, the classifiers do not perform as well as they could. In recent years, deep learning has developed rapid and recent demonstration for the most advanced performance in various fields. And it is a new machine learning algorithm with powerful computing capabilities and overcomes the limitations of traditional machine learning methods such as weak adaptability to data. It can achieve high prediction performance when biology can provide larger and larger dataset. Additionally, deep learning can usually handle high-dimensional, noisy data with non-linear relationships in biology.

6.1 Motivation

An identification method is proposed in this work using protein sequences to identify proteins (genes) that are responsible for having PD. In this paper, a deep learning model, namely the Multi-Layer Perceptron (MLP) and Long Short Term Memory (LSTM) are adopted for PD genes identification. The major contributions of this paper are as follows.

1. Protein sequences are being used as prior knowledge.
2. Twelve physicochemical properties have been used to extract features.
3. A deep learning method has been proposed using MLP and LSTM for PD genes identification.
4. A comparative study with existing systems is carried out to show the effectiveness

of our proposed model.

6.2 Problem Statement

The proposed system model for identifying Parkinson's disease genes has been defined in this section as shown in Figure 6.1. The proposed method comprises of three steps: (1) Utilize physicochemical properties of amino acids to translate protein sequences into numerical features; (2) Stepwise Forward Selection and Backward Elimination method (FSBE) is used to extract best features and remove the worst from remaining attributes; (3) train the models. The steps to identify Parkinson's disease genes using MLP and LSTM is described in Algorithm 6.1.

6.2.1 Extracting features from protein sequences

Extracting features for both disease and unknown genes is one of the most significant tasks in identifying disease genes. In this work, we use protein sequences to characterize genes and used three representation methods to extract information encoded in proteins, such as Normalized Moreau–Broto Autocorrelation (NA) [24], Geary Autocorrelation (GA) [25] and Moran Autocorrelation (MA) [26]. These methods represent the neighboring impact between amino acids with a specific number of amino acids separated in their sequence using their specific physicochemical property. Also, make it possible to find patterns that run through whole sequence. The reason using these methods is to avoid losing important information hidden in the protein sequences. Moreover, these methods are being used in several other works [22] also and have advantage over other methods. These representation methods used their physicochemical properties of amino acids to explain the neighboring effect between amino acids with other amino acids within a sequence. These methods help to gain relevant information, which is unknown in protein sequences. Since all the representation methods are established on physicochemical properties, hence we used twelve physicochemical properties in this paper to acquire further knowledge regarding amino acid sequences. We have utilized twelve physicochemical properties to provide more knowledge about amino acid sequence. The physicochemical properties used are polarity [110], residue accessible surface area in tripeptide [111], hydrophilicity [112], polarizability [113], solvation free energy [114], entropy of formation [115], partition coefficient [116], amino acid composition (AAC) [110], hydrophobicity [117], transfer-free energy [118], CC in Regression analysis [119], and graph shape index [120] respectively. These properties are utilized to obtain the features. Min-max normalization method is applied to normalize the physicochemical properties as shown in Eq. 6.1.

Algorithm 6.1 Algorithm to identify Parkinson's disease genes using MLP and LSTM.

Start

Input : Protein sequences (genes)
Identify : gene is PD or not using feature vector
Output : S={0,1}

initialization
foreach protein sequence **do**
- extractGeary()

$$GA(l) = \frac{\frac{1}{2(N-1)}\Sigma_{i=1}^{N-1}(P_i-\rho_{i+l})^2}{\frac{1}{N-1}\Sigma_{i=1}^{N}(P_i-\tilde{P}')^2}$$

- extractMoran()

$$MA(l) = \frac{\frac{1}{N-1}\Sigma_{i=1}^{N-1}(P_i-\tilde{P}')(P_{i+1}-\tilde{P}')}{\frac{1}{N-1}\Sigma_{i=1}^{N}(P_i-\tilde{P}')^2}$$

- extractMoreauBroto()

$$AC(l) = \Sigma_{i=1}^{N-1} P_i P_{i+1}$$

$$NAC(l) = \frac{AC(l)}{N-1}$$

end
form : feature factor
foreach method **do**
forward selection
backward elimination
end

for numeric data n **do**
LSTM_train(n)
foreach Params, set do
epochs = 80
fully connected layers = 2
hidden units = 150
learning rate = 0.01
end
MLP_train(n)
foreach Params, set do
epochs = 80
fully connected layers = 2
hidden units = 100
end
Calculate Precision, Recall
Calculate F-score
LSTM_test<-Predict{0,1}
Accuracy = accuracy, V accuracy
MLP_test<-Predict{0,1}
Accuracy = accuracy, V accuracy

End

$$P_{xy} = \frac{P_{xy} - P_{y,min}}{P_{y,max} - P_{y,min}} \tag{6.1}$$

Where P_{xy} represents y-th descriptor value for x-th amino acid, $P_{y,min}$ is y-th descriptor minimum value of amino acids and $P_{y,max}$ is y-th descriptor maximum value of amino acids.

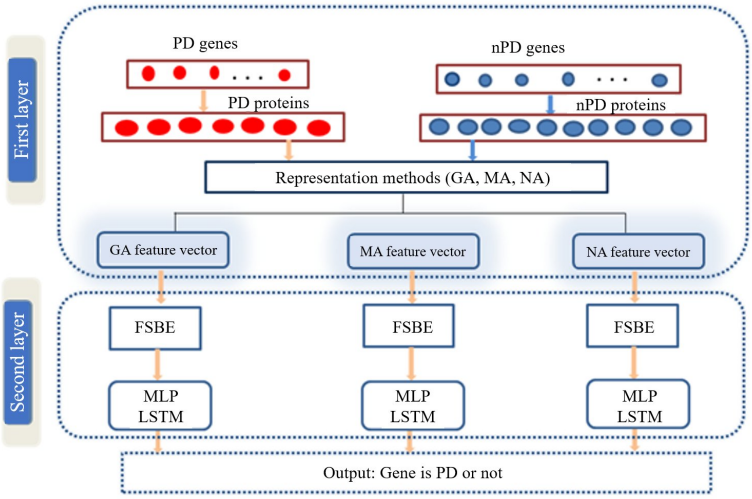


Figure 6.1: Proposed system model for PD identification

6.3 Experimental setup

The proposed Parkinson's disease gene identification method has been constructed using two separate methods, one is MLP and other one is LSTM. The Keras [123] library has been used for implementing deep learning methods due to its user-friendly nature. Table 6.1 shows the different optimization factors that have been organized for training and testing the proposed model. The MLP method has been constructed with three hidden layers, while LSTM with a single hidden layer. The initial weights are uniform between different layers for both the models. The output layer uses the sigmoid activation function to predict probability of having a disease gene or not. Adam optimization method to update network weights is applied to both the models. Both models use the multi-class logarithmic function of cross entropy as the loss function. The number of training examples applied to the input layer before the weight update is 320 (batch size). The early stopping method [124] is used to determine the number of training epochs. Initially 100 epochs are taken to analyse the performance of model. It has been observed that after 80 epochs, the model performance stopped showing any significant improvement because the loss and accuracy value remains almost same till 80 epochs. So, our models

Table 6.1: Parameters used for MLP and LSTM

Tuning Parameters	MLP	LSTM
Model Initialization		
Hidden Layers	3	1
Hidden Units	100, 70 and 50	Gated memory units 128
Activation function	ReLU, tanh, Sigmoid	ReLU, tanh, Sigmoid
Layer type	Dense	
Dropout		
Model compilation		
Loss function	Categorical cross-entropy	
Optimizer	Adam	
Model training		
Batch Size	320	
Epoch	80	

are trained for 80 epochs. The next section discusses the results of the proposed model using configured parameters.

6.4 Results and discussion

The performance of MLP and LSTM methods has been studied on the imbalanced dataset in this section. Firstly, the optimal number of features extracted through feature selection and backward elimination method has been reviewed and optimised. Then, the influence of sequence representation methods has been evaluated on the performance of both MLP and LSTM methods. Finally, a comparison between our proposed method and other disease gene identification methods has been done to obtain relatively negative data to confirm the effectiveness of method.

6.4.1 Experimental data

The dataset is formulated by obtaining human Parkinson's (PD) and non-Parkinson's (nPD) protein sequences (genes) from NCBI [23], Ensembl [24] and UNIPROT [25] databases. The extracted sequences were saved as fasta file. The obtained dataset was then cleaned by eliminating duplicate and partial protein sequences. Total numbers of sequences are 2815, in which 1220 are PD (positive data) and 1595 are nPD (negative data). Each sequence is then transformed into three feature vectors using GA, MA, NA sequence representation methods. This task was performed using protr package in R.

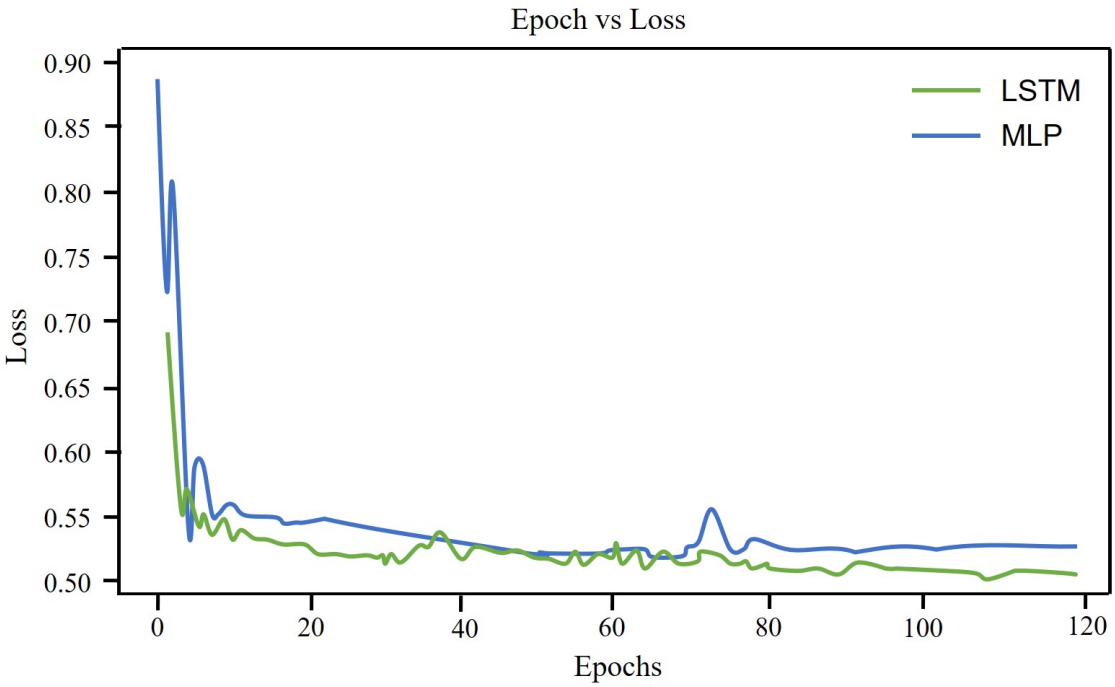


Figure 6.2: Epoch vs loss for MLP and LSTM methods

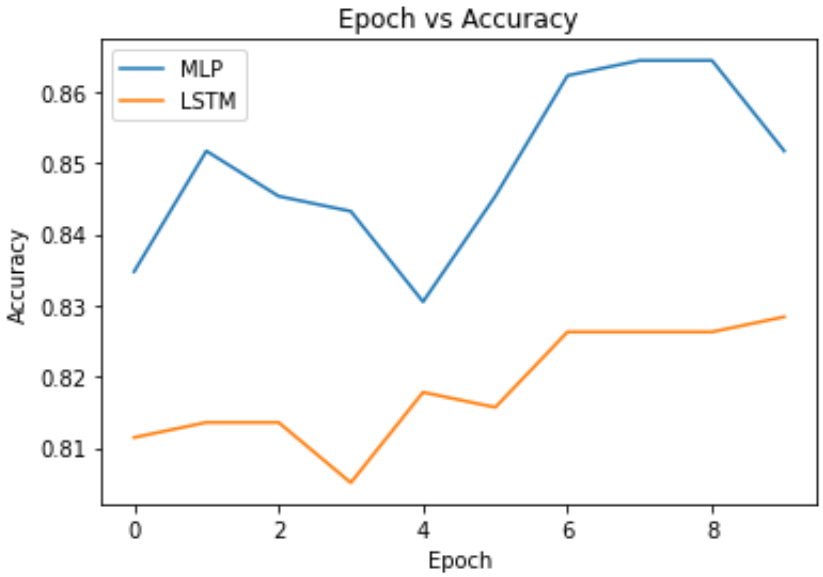


Figure 6.3: Epoch vs Accuracy for MLP and LSTM methods

6.4.2 Performance of the proposed method

To evaluate the robustness of methods, we have endeavoured to train and test the model using both LSTM and MLP based on NA, GA and MA representation methods individually. The outcomes of these methods are displayed in two ways. At first, we evaluate the results after using all the features. Secondly, we use feature reduction methods to

evaluate the results. Figure 6.4 and Figure 6.6 shows the sequence representation method using all features respectively. Figure 6.5 and Figure 6.7 indicates the results of each representation method after applying the feature reduction method. It can be found that the performance of the GA method is better than the performance of the other ones. Moreover, Figure 6.5 and Table 6.2 indicates the positive effect of using feature selection for LSTM in improving the experimental results. Figure 6.7 and Table 6.2 indicates the positive effect of using feature selection for MLP in improving the experimental results. We used an imbalanced data set to study the performance of this method, as the number of unknown genes is more than that of disease genes. It can be found that MLP performs extremely well even with a small amount of data, and when the amount of data is large, LSTM model performs better [125]. To prevent over fitting, we applied the early stopping technique. In this technique, as long as the accuracy of the training set is improving, the network will stop learning and the accuracy of the validation set will stop improving or decreasing. Figure 6.2 shows the epoch versus loss graph for both MLP and LSTM methods. As it is shown that loss stops decreasing after 80 epochs. So, stopping criteria will be activated at 80th epoch. Figure 6.3 shows the epoch versus accuracy graph for both MLP and LSTM methods. It can be concluded that MLP performs better than LSTM.

6.4.3 Discussion

We have compared the F-Score of proposed MLP and LSTM methods on imbalanced dataset. The results obtained from different representation methods for both MLP and LSTM is shown in Table 6.2. The outcomes of these methods are displayed in two ways. At first, we evaluate the results after using all the features. Secondly, we have used feature reduction methods to evaluate the results.

We used an imbalanced data set to study the performance of this method, as the number of unknown genes is more than that of disease genes. Figures 6.4 - 6.7 shows the performances without and with feature selection for both MLP and LSTM methods. It can be observed from the plots that method with reduced features show better performance than methods without feature reduction. For example, the F-Score of GA for MLP is improved from 83.4% to 85%, while for LSTM F-Score is improved from 81.45% to 83.96%. Also, it has been observed that performance of GA is superior to other representation methods. We can say that MLP performs extremely well even with a small amount of data, and when the amount of data is large, LSTM model performs better. The proposed MLP method for GA representation with reduced features significantly outperforms LSTM method and produces stable and scalable performance.

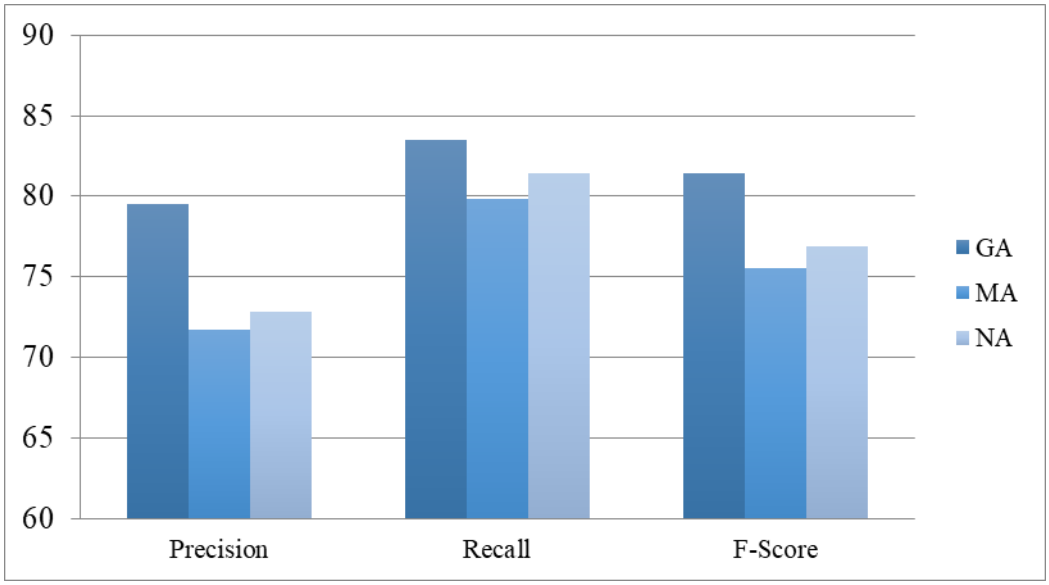


Figure 6.4: Performance comparison(%) of representation methods for LSTM

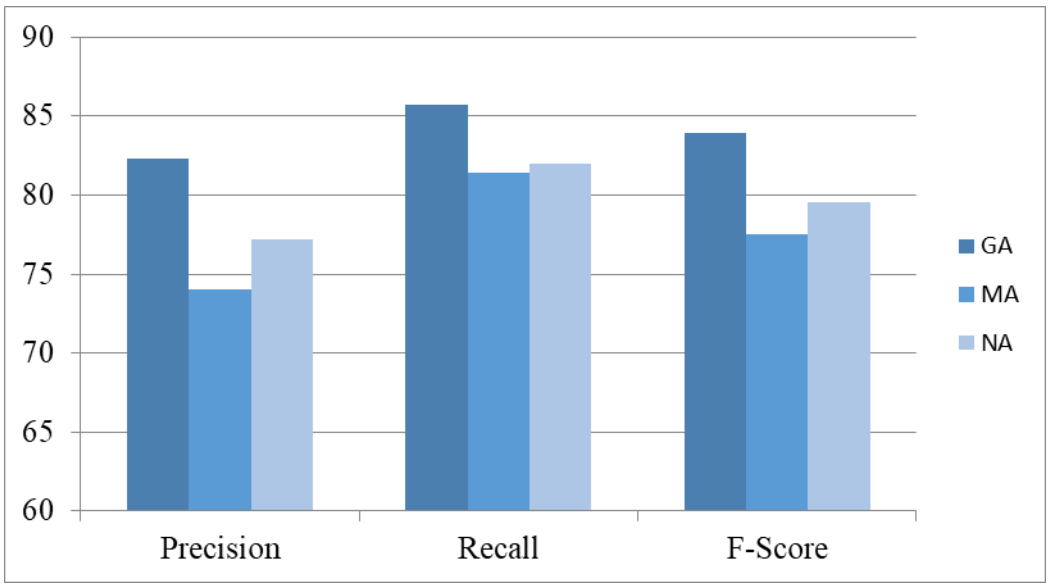


Figure 6.5: Performance (percentage) of representation methods using feature selection for LSTM

Based on the results obtained, we conclude that multi-layer perceptron method has higher accuracy in the diagnosis and prediction of neurological diseases, and is superior to other classifiers. Due to the complexity of genetic and microarray data, it is difficult to make an accurate diagnosis, so computer-aided advanced machine learning technology is used to improve the prediction accuracy and treatment level of neurological diseases.

In this work, we used feature selection and backward elimination method in data preprocessing and then applied deep learning methods on reduced features for gene iden-

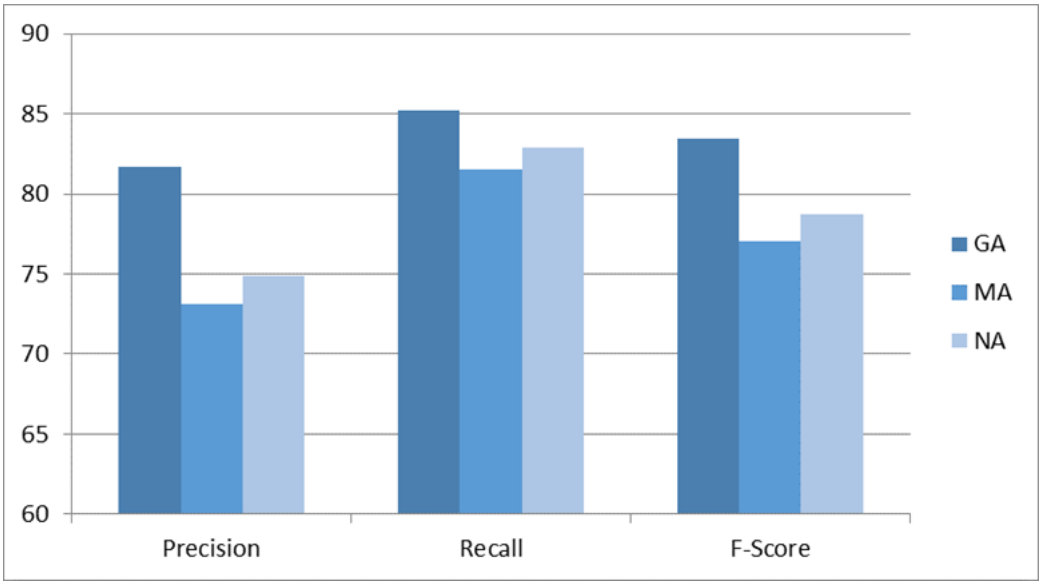


Figure 6.6: Performance (percentage) of representation methods for MLP

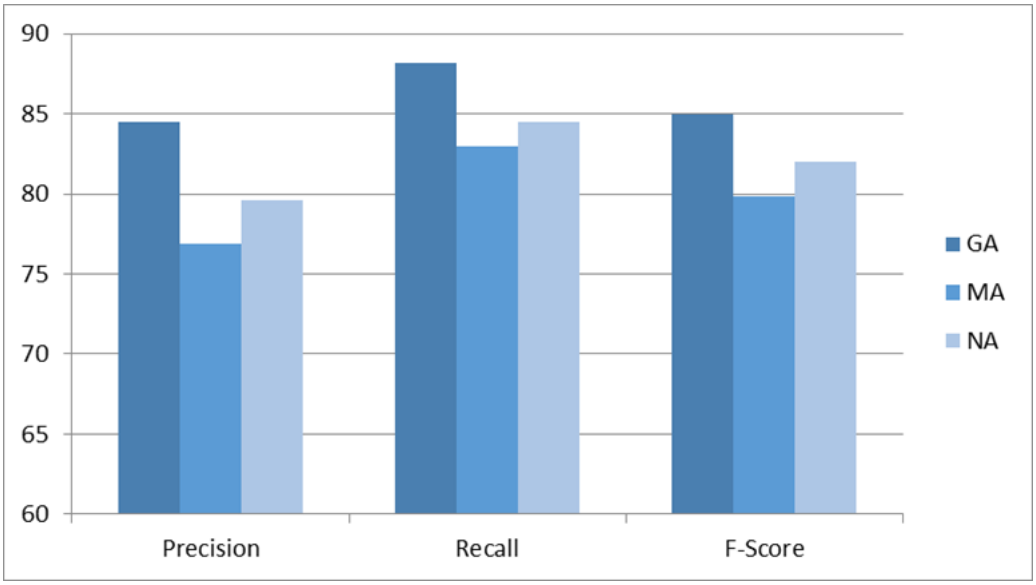


Figure 6.7: Performance (percentage) of representation methods using feature selection for MLP

tification. According to the results obtained from previous research on different data sets, the performance of deep learning methods is more advanced and better than other traditional machine learning classifiers. This motivation prompted us to propose deep learning model, which produces unbiased and stable classification model. Additionally, Deep learning can effectively discover patterns in high-dimensional data. The results obtained shows that MLP method has greater performance where feature selection is done with FSFE method.

Table 6.2: Performances of sequence representation methods with and without feature selection methods

Methods	No. of features	Precision	Recall	F-Score
Without feature selection for LSTM				
GA	360	79.5	83.5	81.45
MA	360	71.7	79.8	75.53
NA	360	72.8	81.4	76.86
With extracted features for LSTM				
GA	65	82.3	85.7	83.96
MA	60	74	81.4	77.52
NA	71	77.2	82.0	79.52
Without feature selection for MLP				
GA	360	81.7	85.2	83.41
MA	360	73.1	81.5	77.07
NA	360	74.9	82.9	78.69
With extracted features for MLP				
GA	65	84.5	88.2	85.0
MA	60	76.9	83	79.83
NA	71	79.6	84.5	81.97

6.4.4 Comparison with existing works

It has been observed from Table 6.3 that proposed MLP with feature reduction method outperforms other state-of-art methods. The proposed method is compared with six state-of-art methods, including, SFM method [48], EPU [10], PUDI [31], ProDiGe [32], Smalter's [27], and Xu's [38]. The protein sequences for both PD and non-PD genes have been collected from NCBI, Ensembl and Uniprot databases. It has been observed that in terms of F-score proposed MLP method on average, is 5.4%, 6.4%, 10.1%, 16.9%, 22.8% and 23.4% higher than EPU, SFM, PUDI, ProDiGe, Smalter's method, Xu's method respectively for imbalanced datasets. The basic difference between the mentioned methods and our proposed method is the prior knowledge used to extract feature vector. In this work, the sequence of proteins was realized as the most common knowledge while in previous work; prior knowledge was affected by noise. The second issue is about classification algorithm used to identify disease genes. Since our preferred sample is unbalanced, we used a precision-recall (PR) curve to handle highly skewed data. In order to establish the relationship between precision and recall and measure the performance of the classifier, the area under the PR curve is preferred. Precision-Recall relationship is shown in Figure 6.8.

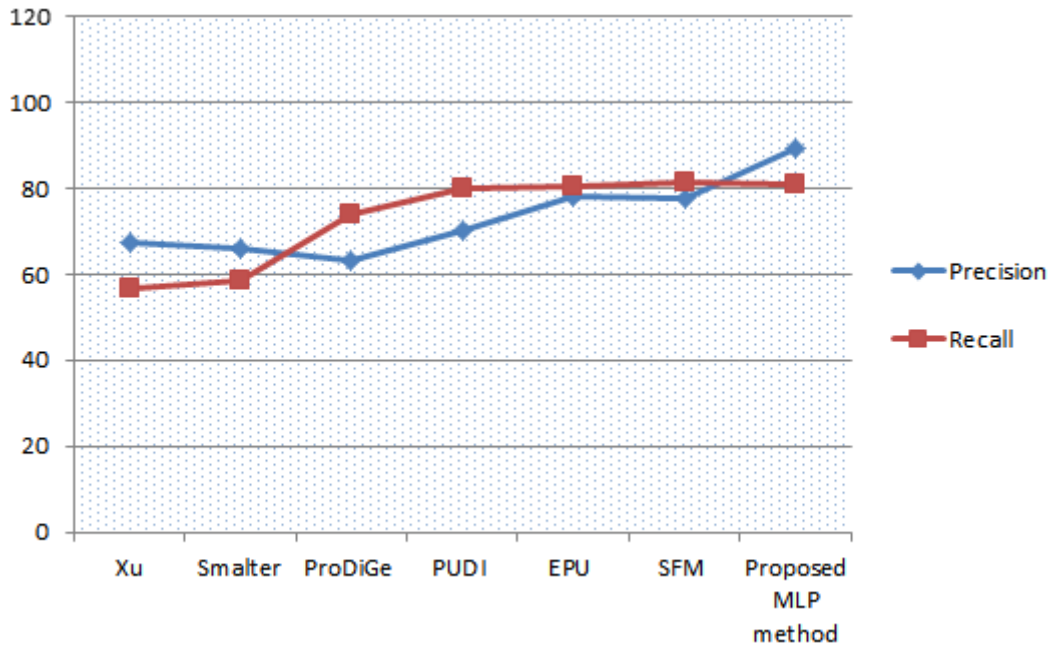


Figure 6.8: Precision-Recall curve for all methods

Table 6.3: Comparative evaluation between proposed and existing systems

Model	**Precision**	**Recall**	**F-Score**
Method	Recall	Precision	F-Score
SFM [23]	81.4	77.9	79.6
EPU [21]	80.4	78.2	78.6
PUDI [20]	80.1	70.3	74.9
ProDiGe [19]	74	63.1	68.1
Smalter's method [16]	58.7	66.2	62.2
Xu's method [15]	56.8	67.4	61.6
Proposed MLP method	88.2	84.5	85

6.5 Summary

In this work,we trained and tested our Parkinson's disease genes dataset on deep learning methods such as MLP and LSTM. We applied Positive Unlabeled learning method to utilize the extracted positive and negative samples as training data. Also, used Feature Selection and Backward Elimination (FSBE) method to extract optimal number of features. It has been found that MLP performs extremely well even with a small amount of data, but when the amount of data is large, LSTM model performs better. The proposed MLP method for GA representation with reduced features significantly outperforms LSTM method and produces stable and scalable performance. Based on the results obtained, we conclude that multi-layer perceptron method has higher accuracy in

the diagnosis and prediction of neurological diseases, and is superior to other classifiers. Due to the complexity of genetic and microarray data, it is difficult to make an accurate diagnosis, so computer-aided advanced machine learning technology is used to improve the prediction accuracy and treatment level of neurological diseases.

The explainability methods to open these black boxes have also been proposed in the literature with human agent interaction [126] [127] [128]. This is the next step for the successful deployment of AI technologies in the medical field to increase the trust and understandability of computational methods in disease detection which are currently out of the scope of this work and would be considered in the future.

Chapter 7

Conclusions and scope for future work

This chapter concludes the research work presented in this thesis and the future scope of the work has also been discussed.

7.1 Conclusion

In this thesis, different approaches have been identified to detect Parkinson's disease genes using protein sequences. The main focus has been given on the multi-voting ensemble model, Positive Unlabeled learning and N-semble learning for Parkinson's disease gene classification. A comprehensive literature review for disease gene identification over machine learning based methods has been performed. Also, various available datasets for gene identification were identified and studied. However, the previous studies differ in the data type used to generate feature vectors, such as gene expression profiles, protein-protein interactions (PPI), protein and biological functions. All of these techniques, though, rely on information about proteins that is achieved from protein domains, gene ontology, and PPI data. Since information is incomplete, time consuming, and noisy, so might not be implemented correctly. Protein sequences are available for all proteins and help solve problems like protein-protein interactions, functional and structural classifications, etc. Therefore, we proposed methods to identify disease genes using protein sequences only.

Firstly, a multi-layer ensemble method has been proposed in this work to identify Parkinson's disease genes. Different machine learning models are selected and trained to build the ensemble method. Hydrophobicity and amino acid properties are used to extract features from protein sequences. Pearson correlation coefficient is used to extract the relevant and best features from high-dimensional data.Machine learning approaches, such as Naive Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Neural Network (nnet), Logistic Regression, Decision Tree (DT), Random Forest (RF) and ensemble methods are compared and analyzed on the basis of performance measures to effectively diagnose PD. After analyzing the classifiers performance measures, the focus was to build a learning algorithm by combining base estimators to get optimal predictive

model. So, an ideal multi-level voting model has been proposed, which integrates various ML models based on their FP rates to retrieve a new voting classifier to retrieve better prediction analysis. The testing benchmark and results of the proposed multi-voting ensemble model with various state-of-the-art methods have been compared.

To overcome the limitation of binary classification, we have proposed second approach to improve the efficiency of method. For second approach, we have applied Positive-Unlabeled based learning to proposed N-semble method for identifying disease genes. Geary Autocorrelation (GA), Moran Autocorrelation (MA) and Normalized Moreau Broto Autocorrelation (NA) representation methods with twelve physicochemical properties of the amino acids are adopted to convert protein sequences into numerical feature vectors. Consequently, t-SNE is applied to extract relevant features. Various machine learning models, includes, Support Vector Machine (SVM), Random Forest (RF), Adaboost, Decision Tree (DT), Xgboost and Gradient Descent are trained to develop an N-semble method. An n-semble method trained a neural network in a special way and integrated three best classification methods based on their F-Score to ensemble the predictions and to achieve more accurate predictive analysis. A comparative study is performed with the existing ensemble and state-of-art methods to show the effectiveness of the proposed model. A case study has been performed to identify novel disease genes for Parkinson's diseases. Finally, we have tested and trained Artificial Neural network approaches on Parkinson's disease genes dataset. MLP and LSTM models are used to classify Parkinson's disease genes. The comparison analysis is also performed with various state-of-the-art techniques, which shows the efficiency of the proposed systems. We have found that the Geary Autocorrelation representation method is characterized by a higher success rate than other representation methods.

In the future, we would consider using a single GA feature vector to combine multiple different classifiers to improve classification results. It has also been analysed that the results achieved by the proposed systems are promising and the methods can be extended to the other neurological diseases as well.

7.2 Future scope

The proposed systems can be used to diagnose the Parkinson's disease genes. Here we present some possible directions for future study in the field of disease gene identification.

1. In future, more biological resources such as gene ontology and gene expression data can be integrated with protein sequences.

2. We may explore different Positive Unlabeled learning methods to extract negative samples from unlabeled dataset and to address the issue of imbalance data.

3. The proposed multi-level ensemble model can be trained on other neurological diseases for disease gene identification.

4. We can also consider how to generate the rules for disease gene identification.

5. To build a robust classifier, an important next step in our algorithm is to further extract the likely positive samples (LP) and the likely negative samples (LN) from genes in the unknown sample US-r which are near the positive and negative classification boundary.

References

[1] Martin Oti and Han G Brunner. The modular nature of genetic diseases. *Clinical genetics*, 71(1):1–11, 2007.

[2] Yana Bromberg. Chapter 15: disease gene prioritization. *PLoS computational biology*, 9(4):e1002902, 2013.

[3] Nivedhitha Mahendran and Durai Raj Vincent PM. A deep learning framework with an embedded-based feature selection approach for the early detection of the alzheimer's disease. *Computers in Biology and Medicine*, 141:105056, 2022.

[4] Fiona J Charlson, Amanda J Baxter, Tarun Dua, Louisa Degenhardt, Harvey A Whiteford, and Theo Vos. Excess mortality from mental, neurological, and substance use disorders in the global burden of disease study 2010. *Mental, Neurological, and Substance Use Disorders*, page 41, 2016.

[5] Hemanta Kumar Nayak, Mrudul Kumar Daga, Rakshit Kumar, Sandeep Kumar Garg, Naresh Kumar, and Pankaj Kumar Mohanty. A series report of autoimmune hypothyroidism associated with hashimoto's encephalopathy: An under diagnosed clinical entity with good prognosis. *Case Reports*, 2010:bcr0120102630, 2010.

[6] Kuljeet Singh Anand and Vikas Dhikav. Hippocampus in health and disease: An overview. *Annals of Indian Academy of Neurology*, 15(4):239, 2012.

[7] Saba Khanam and Yasir H Siddique. Dopamine: agonists and neurodegenerative disorders. *Current drug targets*, 19(14):1599–1611, 2018.

[8] J William Langston. Parkinson's disease: current and future challenges. *Neurotoxicology*, 23(4-5):443–450, 2002.

[9] Abdulaziz Yousef and Nasrollah Moghadam Charkari. A novel method based on new adaptive lvq neural network for predicting protein–protein interactions from protein sequences. *Journal of theoretical biology*, 336:231–239, 2013.

[10] Peng Yang, Xiaoli Li, Hon-Nian Chua, Chee-Keong Kwoh, and See-Kiong Ng. Ensemble positive unlabeled learning for disease gene identification. *PloS one*, 9(5):e97079, 2014.

[11] Peng Han, Peng Yang, Peilin Zhao, Shuo Shang, Yong Liu, Jiayu Zhou, Xin Gao, and Panos Kalnis. Gcn-mf: disease-gene association identification by graph convolutional networks and matrix factorization. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 705–713, 2019.

[12] Han G Brunner and Marc A Van Driel. From syndrome families to functional

genomics. *Nature Reviews Genetics*, 5(7):545–551, 2004.

[13] Euan A Adie, Richard R Adams, Kathryn L Evans, David J Porteous, and Ben S Pickard. Suspects: enabling fast and effective prioritization of positional candidates. *Bioinformatics*, 22(6):773–774, 2006.

[14] Frances S Turner, Daniel R Clutterbuck, and Colin AM Semple. Pocus: mining genomic sequence annotation to predict disease genes. *Genome biology*, 4(11):1–9, 2003.

[15] Jing Chen, Huan Xu, Bruce J Aronow, and Anil G Jegga. Improved human disease candidate gene prioritization using mouse phenotype. *BMC bioinformatics*, 8(1):1–13, 2007.

[16] Stein Aerts, Diether Lambrechts, Sunit Maity, Peter Van Loo, Bert Coessens, Frederik De Smet, Leon-Charles Tranchevent, Bart De Moor, Peter Marynen, Bassem Hassan, et al. Gene prioritization through genomic data fusion. *Nature biotechnology*, 24(5):537–544, 2006.

[17] Wangshu Zhang, Fengzhu Sun, and Rui Jiang. Integrating multiple protein-protein interaction networks to prioritize disease genes: a bayesian regression approach. *BMC bioinformatics*, 12(1):1–10, 2011.

[18] Peng Yang, Xiaoli Li, Min Wu, Chee-Keong Kwoh, and See-Kiong Ng. Inferring gene-phenotype associations via global protein complex network propagation. *PloS one*, 6(7):e21502, 2011.

[19] Abdulaziz Yousef and Nasrollah Moghadam Charkari. A novel method based on physicochemical properties of amino acids and one class classification algorithm for disease gene identification. *Journal of biomedical informatics*, 56:300–306, 2015.

[20] Ugo Ala, Rosario Michael Piro, Elena Grassi, Christian Damasco, Lorenzo Silengo, Martin Oti, Paolo Provero, and Ferdinando Di Cunto. Prediction of human disease genes by human-mouse conserved coexpression analysis. *PLoS computational biology*, 4(3):e1000043, 2008.

[21] Jan Freudenberg and P Propping. A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics*, 18(suppl_2):S110–S115, 2002.

[22] Tijl De Bie, Leon-Charles Tranchevent, Liesbeth MM Van Oeffelen, and Yves Moreau. Kernel-based data fusion for gene prioritization. *Bioinformatics*, 23(13):i125–i132, 2007.

[23] NCBI Database. Ncbi database. `https://instr.iastate.libguides.com/NCBI`. [Online; accessed 2022, August 25].

[24] Ensembl Database. Ensemble genome database. `https://ensemblgenomes.org`. [Online; accessed 2022, August 25].

[25] Uniprot Database. Protein information and functional information uniprot database. `https://www.uniprot.org`. [Online; accessed 2022, August 25].

[26] Euan A Adie, Richard R Adams, Kathryn L Evans, David J Porteous, and Ben S Pickard. Speeding disease gene discovery by sequence based candidate prioritization. *BMC bioinformatics*, 6(1):1–13, 2005.

[27] Aaron Smalter, Seak Fei Lei, and Xue-wen Chen. Human disease-gene classification with integrative sequence-based and topological features of protein-protein interaction networks. In *2007 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2007)*, pages 209–216. IEEE, 2007.

[28] Predrag Radivojac, Kang Peng, Wyatt T Clark, Brandon J Peters, Amrita Mohan, Sean M Boyle, and Sean D Mooney. An integrated approach to inferring gene–disease associations in humans. *Proteins: Structure, Function, and Bioinformatics*, 72(3):1030–1037, 2008.

[29] SHIVAKUMAR Keerthikumar, SAHELY Bhadra, KUMARAN Kandasamy, RAJESH Raju, YL Ramachandra, CHIRANJIB Bhattacharyya, KOHSUKE Imai, OSAMU Ohara, SUJATHA Mohan, and AKHILESH Pandey. Prediction of candidate primary immunodeficiency disease genes using a support vector machine learning approach. *DNA research*, 16(6):345–351, 2009.

[30] Jiabao Sun, Jagdish C Patra, and Yongjin Li. Functional link artificial neural network-based disease gene prediction. In *2009 International Joint Conference on Neural Networks*, pages 3003–3010. IEEE, 2009.

[31] Peng Yang, Xiao-Li Li, Jian-Ping Mei, Chee-Keong Kwoh, and See-Kiong Ng. Positive-unlabeled learning for disease gene identification. *Bioinformatics*, 28(20):2640–2647, 2012.

[32] Fantine Mordelet and Jean-Philippe Vert. Prodige: Prioritization of disease genes with multitask machine learning from positive and unlabeled examples. *BMC bioinformatics*, 12(1):1–15, 2011.

[33] Duc-Hau Le, Nguyen Xuan Hoai, and Yung-Keun Kwon. A comparative study of classification-based machine learning methods for novel disease gene prediction. In *Knowledge and Systems Engineering*, pages 577–588. Springer, 2015.

[34] Ranjan Kumar Barman, Anirban Mukhopadhyay, Ujjwal Maulik, and Santasabuj Das. Identification of infectious disease-associated host genes using machine learning techniques. *BMC bioinformatics*, 20(1):1–12, 2019.

[35] Jiajie Peng, Jiaojiao Guan, and Xuequn Shang. Predicting parkinson's disease genes based on node2vec and autoencoder. *Frontiers in genetics*, 10:226, 2019.

[36] Akram Vasighizaker, Alok Sharma, and Abdollah Dehzangi. A novel one-class classification approach to accurately predict disease-gene association in acute myeloid

leukemia cancer. *PloS one*, 14(12):e0226115, 2019.

[37] Duc-Hau Le and Manh-Hien Nguyen. Towards more realistic machine learning techniques for prediction of disease-associated genes. In *Proceedings of the Sixth International Symposium on Information and Communication Technology*, pages 116–120, 2015.

[38] Jianzhen Xu and Yongjin Li. Discovering disease-genes by topological features in human protein–protein interaction network. *Bioinformatics*, 22(22):2800–2805, 2006.

[39] Yun Xiao, Chaohan Xu, Yanyan Ping, Jinxia Guan, Huihui Fan, Yiqun Li, and Xia Li. Differential expression pattern-based prioritization of candidate genes through integrating disease-specific expression data. *Genomics*, 98(1):64–71, 2011.

[40] Wook-Yeon Hwang. Biological feature selection and disease gene identification using new stepwise random forests. *Industrial Engineering and Management Systems*, 16(1):64–79, 2017.

[41] Apichat Suratanee and Kitiporn Plaimas. Network-based association analysis to infer new disease-gene relationships using large-scale protein interactions. *PLoS One*, 13(6):e0199435, 2018.

[42] Walid A Misba, Mark Lozano, Damien Querlioz, and Jayasimha Atulasimha. Energy efficient learning with low resolution stochastic domain wall synapse based deep neural networks. *arXiv preprint arXiv:2111.07284*, 2021.

[43] Gholam-Hossein Jowkar and Eghbal G Mansoori. Perceptron ensemble of graph-based positive-unlabeled learning for disease gene identification. *Computational biology and chemistry*, 64:263–270, 2016.

[44] Fabien Letouzey, François Denis, and Rémi Gilleron. Learning from positive and unlabeled examples. In *International Conference on Algorithmic Learning Theory*, pages 71–85. Springer, 2000.

[45] François Denis, Rémi Gilleron, and Fabien Letouzey. Learning from positive and unlabeled examples. *Theoretical Computer Science*, 348(1):70–83, 2005.

[46] Holger Prokisch, Christophe Andreoli, U Ahting, K Heiss, Andreas Ruepp, Curt Scharfe, and Thomas Meitinger. Mitop2: the mitochondrial proteome database—now including mouse data. *Nucleic acids research*, 34(suppl_1):D705–D711, 2006.

[47] Thanh-Phuong Nguyen and Tu-Bao Ho. Detecting disease genes based on semi-supervised learning and protein–protein interaction networks. *Artificial intelligence in medicine*, 54(1):63–71, 2012.

[48] Abdulaziz Yousef and Nasrollah Moghadam Charkari. Sfm: a novel sequence-based fusion method for disease genes identification and prioritization. *Journal of*

theoretical biology, 383:12–19, 2015.

[49] Xingyu Chen, Qixing Huang, Yang Wang, Jinlong Li, Haiyan Liu, Yun Xie, Zong Dai, Xiaoyong Zou, and Zhanchao Li. A deep learning approach to identify association of disease–gene using information of disease symptoms and protein sequences. *Analytical Methods*, 12(15):2016–2026, 2020.

[50] Ritu Gautam and Manik Sharma. Prevalence and diagnosis of neurological disorders using different deep learning techniques: a meta-analysis. *Journal of medical systems*, 44(2):1–24, 2020.

[51] Ping Luo, Yuanyuan Li, Li-Ping Tian, and Fang-Xiang Wu. Enhancing the prediction of disease–gene associations with multimodal deep learning. *Bioinformatics*, 35(19):3735–3742, 2019.

[52] Mohamad Koohi-Moghadam, Haibo Wang, Yuchuan Wang, Xinming Yang, Hongyan Li, Junwen Wang, and Hongzhe Sun. Predicting disease-associated mutation of metal-binding sites in proteins using a deep learning approach. *Nature Machine Intelligence*, 1(12):561–567, 2019.

[53] Padideh Danaee, Reza Ghaeini, and David A Hendrix. A deep learning approach for cancer detection and relevant gene identification. In *Pacific symposium on biocomputing 2017*, pages 219–229. World Scientific, 2017.

[54] Nazanin Fouladgar, Marjan Alirezaie, and Kary Främling. Cn-waterfall: a deep convolutional neural network for multimodal physiological affect detection. *Neural Computing and Applications*, 34(3):2157–2176, 2022.

[55] Nazanin Fouladgar, Marjan Alirezaie, Kary Främling, et al. Cn-waterfall. 2021.

[56] Nazanin Fouladgar, Marjan Alirezaie, and Kary Främling. Metrics and evaluations of time series explanations: An application in affect computing. *IEEE Access*, 10:23995–24009, 2022.

[57] Sumit Tripathi, Ashish Verma, and Neeraj Sharma. Automatic segmentation of brain tumour in mr images using an enhanced deep learning approach. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 9(2):121–130, 2021.

[58] Pankaj K Jain, Neeraj Sharma, Luca Saba, Kosmas I Paraskevas, Mandeep K Kalra, Amer Johri, John R Laird, Andrew N Nicolaides, and Jasjit S Suri. Unseen artificial intelligence—deep learning paradigm for segmentation of low atherosclerotic plaque in carotid ultrasound: a multicenter cardiovascular study. *Diagnostics*, 11(12):2257, 2021.

[59] Ying Cui, Meng Cai, Yang Dai, and H Eugene Stanley. A hybrid network-based method for the detection of disease-related genes. *Physica A: Statistical Mechanics and Its Applications*, 492:389–394, 2018.

[60] Ibrahim Ibrahim and Adnan Abdulazeez. The role of machine learning algorithms for diagnosing diseases. *Journal of Applied Science and Technology Trends*, 2(01):10–19, 2021.

[61] Jiawei Han, Jian Pei, and Hanghang Tong. *Data mining: concepts and techniques*. Morgan kaufmann, 2022.

[62] Chun-Ling Chuang. Case-based reasoning support for liver disease diagnosis. *Artificial Intelligence in Medicine*, 53(1):15–23, 2011.

[63] Saeed Abu-Nimeh, Dario Nappa, Xinlei Wang, and Suku Nair. A comparison of machine learning techniques for phishing detection. In *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit*, pages 60–69, 2007.

[64] Mohammad M Ghiasi, Sohrab Zendehboudi, and Ali Asghar Mohsenipour. Decision tree-based diagnosis of coronary artery disease: Cart model. *Computer methods and programs in biomedicine*, 192:105400, 2020.

[65] V Pergialiotis, A Pouliakis, C Parthenis, V Damaskou, C Chrelias, N Papantoniou, and I Panayiotides. The utility of artificial neural networks and classification and regression trees for the prediction of endometrial cancer in postmenopausal women. *Public Health*, 164:1–6, 2018.

[66] Paraskevas Tsangaratos and Ioanna Ilia. Comparison of a logistic regression and naïve bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size. *Catena*, 145:164–179, 2016.

[67] VR Balaji, ST Suganthi, R Rajadevi, V Krishna Kumar, B Saravana Balaji, and Sanjeevi Pandiyan. Skin disease detection and segmentation using dynamic graph cut algorithm and classification through naive bayes classifier. *Measurement*, 163:107922, 2020.

[68] Yudong Zhang, Preetha Phillips, Shuihua Wang, Genlin Ji, Jiquan Yang, and Jianguo Wu. Fruit classification by biogeography-based optimization and feedforward neural network. *Expert Systems*, 33(3):239–253, 2016.

[69] Jyoti Prakash Medhi and Samarendra Dandapat. An effective fovea detection and automatic assessment of diabetic maculopathy in color fundus images. *Computers in biology and medicine*, 74:30–44, 2016.

[70] Shuihua Wang, Preetha Phillips, Aijun Liu, and Sidan Du. Tea category identification using computer vision and generalized eigenvalue proximal svm. *Fundamenta Informaticae*, 151(1-4):325–339, 2017.

[71] Zayrit Soumaya, Belhoussine Drissi Taoufiq, Nsiri Benayad, Benba Achraf, and Abdelkrim Ammoumou. A hybrid method for the diagnosis and classifying parkinson's patients based on time–frequency domain properties and k-nearest neighbor. *Journal of Medical Signals and Sensors*, 10(1):60, 2020.

[72] Stephan Dreiseitl and Lucila Ohno-Machado. Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, 35(5-6):352–359, 2002.

[73] Sawinder Kaur, Parteek Kumar, and Ponnurangam Kumaraguru. Automating fake news detection system using multi-level voting model. *Soft Computing*, 24(12):9049–9069, 2020.

[74] Sajida Perveen, Muhammad Shahbaz, Aziz Guergachi, and Karim Keshavjee. Performance analysis of data mining classification techniques to predict diabetes. *Procedia Computer Science*, 82:115–121, 2016.

[75] Chao Tan, Hui Chen, and Chengyun Xia. Early prediction of lung cancer based on the combination of trace element analysis in urine and an adaboost algorithm. *Journal of pharmaceutical and biomedical analysis*, 49(3):746–752, 2009.

[76] Kandala NVPS Rajesh and Ravindra Dhuli. Classification of imbalanced ecg beats using re-sampling techniques and adaboost ensemble classifier. *Biomedical Signal Processing and Control*, 41:242–254, 2018.

[77] Devesh Mishra, KK Agrawal, RS Yadav, and Rekha Srivstava. Design of out pipe crawler for oil refinery based on analysis & classification of locomotion and adhesion techniques. In *2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, pages 1–7. IEEE, 2018.

[78] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407, 2000.

[79] Neeraj Sharma, Lalit M Aggarwal, et al. Automated medical image segmentation techniques. *Journal of medical physics*, 35(1):3, 2010.

[80] Changkui Lei, Jun Deng, Kai Cao, Yang Xiao, Li Ma, Weifeng Wang, Teng Ma, and Chimin Shu. A comparison of random forest and support vector machine approaches to predict coal spontaneous combustion in gob. *Fuel*, 239:297–311, 2019.

[81] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[82] Matthew R Bennett and Marcin Budka. Data analysis and techniques. In *Digital Technology for Forensic Footwear Analysis and Vertebrate Ichnology*, pages 91–135. Springer, 2019.

[83] Abbas Raza Ali and Marcin Budka. An automated approach for timely diagnosis and prognosis of coronavirus disease. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

[84] Hassan Ramchoun, Youssef Ghanou, Mohamed Ettaouil, and Mohammed Amine Janati Idrissi. Multilayer perceptron: Architecture optimization and training. 2016.

[85] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[86] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[87] Thomas G Dietterich. Ensemble learning, the handbook of brain theory and neural networks, ma arbib, 2002.

[88] Carlos Alberto de Araújo Padilha, Dante Augusto Couto Barone, and Adrião Duarte Dória Neto. A multi-level approach using genetic algorithms in an ensemble of least squares support vector machines. *Knowledge-Based Systems*, 106:85–95, 2016.

[89] Robi Polikar. Ensemble based systems in decision making. *IEEE Circuits and systems magazine*, 6(3):21–45, 2006.

[90] Thomas G Dietterich. Machine-learning research. *AI magazine*, 18(4):97–97, 1997.

[91] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[92] Peter Bartlett, Yoav Freund, Wee Sun Lee, and Robert E Schapire. Boosting the margin: A new explanation for the effectiveness of voting methods. *The annals of statistics*, 26(5):1651–1686, 1998.

[93] Magdalena Graczyk, Tadeusz Lasota, Bogdan Trawiński, and Krzysztof Trawiński. Comparison of bagging, boosting and stacking ensembles applied to real estate appraisal. In *Asian conference on intelligent information and database systems*, pages 340–350. Springer, 2010.

[94] Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

[95] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992.

[96] Chao Gao, Hanbo Sun, Tuo Wang, Ming Tang, Nicolaas I Bohnen, Martijn LTM Müller, Talia Herman, Nir Giladi, Alexandr Kalinin, Cathie Spino, et al. Model-based and model-free machine learning techniques for diagnostic prediction and classification of clinical outcomes in parkinson's disease. *Scientific reports*, 8(1):1–21, 2018.

[97] Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701, 1937.

[98] Shinji Sakamoto, Algenti Lala, Tetsuya Oda, Vladi Kolici, Leonard Barolli, and Fatos Xhafa. Analysis of wmn-hc simulation system data using friedman test. In

2015 Ninth international conference on complex, intelligent, and software intensive systems, pages 254–259. IEEE, 2015.

[99] Lei Xu, Guangmin Liang, Changrui Liao, Gin-Den Chen, and Chi-Chang Chang. K-skip-n-gram-rf: a random forest based method for alzheimer's disease protein identification. *Frontiers in genetics*, 10:33, 2019.

[100] Yu Miao, Huiyan Jiang, Huiling Liu, and Yu-dong Yao. An alzheimers disease related genes identification method based on multiple classifier integration. *Computer Methods and Programs in Biomedicine*, 150:107–115, 2017.

[101] Stefan Simm, Jens Einloft, Oliver Mirus, and Enrico Schleiff. 50 years of amino acid hydrophobicity scales: revisiting the capacity for peptide classification. *Biological research*, 49:1–19, 2016.

[102] Oliviero Carugo. Amino acid composition and protein dimension. *Protein Science*, 17(12):2187–2191, 2008.

[103] Luigi Cerulo, Charles Elkan, and Michele Ceccarelli. Learning gene regulatory networks from only positive and unlabeled data. *BMC bioinformatics*, 11(1):1–16, 2010.

[104] Bolan Linghu, Evan S Snitkin, Zhenjun Hu, Yu Xia, and Charles DeLisi. Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. *Genome biology*, 10(9):1–17, 2009.

[105] Chi-Yuan Yu, Lih-Ching Chou, and Darby Tien-Hao Chang. Predicting protein-protein interactions in unbalanced data using the primary structure of proteins. *BMC bioinformatics*, 11(1):1–10, 2010.

[106] Yoshinori Fukasawa, Ross KK Leung, Stephen KW Tsui, and Paul Horton. Evolutionary sequence divergence predicts protein sub-cellular localization signals. In *2011 IEEE International Conference on Systems Biology (ISB)*, pages 307–312. IEEE, 2011.

[107] Zhi-Ping Feng and Chun-Ting Zhang. Prediction of membrane protein types based on the hydrophobic index of amino acids. *Journal of protein chemistry*, 19(4):269–275, 2000.

[108] Jun-Feng Xia, Kyungsook Han, and De-Shuang Huang. Sequence-based prediction of protein-protein interactions by means of rotation forest and autocorrelation descriptor. *Protein and Peptide Letters*, 17(1):137–145, 2010.

[109] Robert R Sokal and Barbara A Thomson. Population structure inferred by local spatial autocorrelation: an example from an amerindian tribal population. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, 129(1):121–131, 2006.

[110] Richard Grantham. Amino acid difference formula to help explain protein evolution. *science*, 185(4154):862–864, 1974.

[111] Cyrus Chothia. The nature of the accessible and buried surfaces in proteins. *Journal of molecular biology*, 105(1):1–12, 1976.

[112] Thomas P Hopp and Kenneth R Woods. Prediction of protein antigenic determinants from amino acid sequences. *Proceedings of the National Academy of Sciences*, 78(6):3824–3828, 1981.

[113] Marvin Charton and Barbara I Charton. The structural dependence of amino acid hydrophobicity parameters. *Journal of theoretical biology*, 99(4):629–644, 1982.

[114] David Eisenberg and Andrew D McLachlan. Solvation energy in protein folding and binding. *Nature*, 319(6050):199–203, 1986.

[115] Cyrus Chothia. One thousand families for the molecular biologist. *Nature*, 357(6379):543–544, 1992.

[116] J Ross Quinlan. Improved use of continuous attributes in c4. 5. *Journal of artificial intelligence research*, 4:77–90, 1996.

[117] Robert M Sweet and David Eisenberg. Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. *Journal of molecular biology*, 171(4):479–488, 1983.

[118] JOEL Janin. Surface and inside volumes in globular proteins. *Nature*, 277(5696):491–492, 1979.

[119] M Prabhakaran and PK Ponnuswamy. Shape and surface features of globular proteins. *Macromolecules*, 15(2):314–320, 1982.

[120] JEAN-LUC FAUCHÈRE, Marvin Charton, Lemont B Kier, Arie Verloop, and Vladimir Pliska. Amino acid side chain parameters for correlation studies in biology and pharmacology. *International journal of peptide and protein research*, 32(4):269–278, 1988.

[121] Wentian Li, Jane E Cerise, Yaning Yang, and Henry Han. Application of t-sne to human genetic data. *Journal of bioinformatics and computational biology*, 15(04):1750017, 2017.

[122] Rishith Rayal, Divya Khanna, Jasminder Kaur Sandhu, Nishtha Hooda, and Prashant Singh Rana. N-semble: neural network based ensemble approach. *International Journal of Machine Learning and Cybernetics*, 10(2):337–345, 2019.

[123] François Chollet et al. keras. github repository. *https://github. com/fchollet/keras¿. Accessed on*, 25:2017, 2015.

[124] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[125] Fang Wu, Jigang Wang, Jiqiang Liu, and Wei Wang. Vulnerability detection with deep learning. In *2017 3rd IEEE international conference on computer and communications (ICCC)*, pages 1298–1302. IEEE, 2017.

[126] Yazan Mualla, Igor Haman Tchappi, Amro Najjar, Timotheus Kampik, Stéphane Galland, and Christophe Nicolle. Human-agent explainability: An experimental case study on the filtering of explanations. In *ICAART (1)*, pages 378–385, 2020.

[127] Yazan Mualla, Igor Tchappi, Timotheus Kampik, Amro Najjar, Davide Calvaresi, Abdeljalil Abbas-Turki, Stéphane Galland, and Christophe Nicolle. A human-agent architecture for explanation formulation.

[128] Amro Najjar, Yazan Mualla, Kamal Singh, and Gauthier Picard. One-to-many multi-agent negotiation and coordination mechanisms to manage user satisfaction. In *the 11th International Workshop on Agent-based Complex Automated Negotiations (ACAN2018)*, 2018.

List of Publications

1. Priya Arora, Ashutosh Mishra, and Avleen Malhi *"Machine learning Ensemble for the Parkinson's disease using protein sequences"*, Multimedia Tools and Applications, Springer, 81:32215–32242, 2022. [SCIE, IF 2.57]
2. Priya Arora, Ashutosh Mishra, and Avleen Malhi *"N-semble-based method for identifying Parkinson's disease genes"*, Neural Computing and Applications, Springer, 2021]. [SCIE, IF 5.1]
3. Priya Arora, Ashutosh Mishra, and Avleen Malhi *"LSTM and MLP based multi-feature extraction for diagnosis of Parkinson's disease genes"*, Health and Technology, Springer. [Under Review]
4. Priya Arora, Ashutosh Mishra, and Avleen Malhi *"An Ensemble machine learning method highlights possible Parkinson's disease genes and accessing performance of re-sampling techniques"*, SN Computer Science, Springer. [Under Review]