Analysis of the Off-line Signature Verification Techniques

Dissertation submitted in partial fulfilment of the requirements for the award of degree of

Master of Engineering

in

Computer Science and Engineering

Submitted by

Gaganpreet Singh

(801532014)

Under the supervision of:

Dr. Singara Singh Kasana

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR UNIVERSITY

PATIALA 147004

July 2017

Certificate

I hereby certify that the matter which is being presented in the dissertation titled, **Analysis of the Off-line Signature Verification Techniques**, in partial fulfilment of the requirements for the award of degree of Master of Engineering in **Computer Science and Engineering** submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work, under the supervision of **Dr. Singara Singh Kasana** and refers other researchers work which is duly listed in the reference section. The matter presented in this dissertation has not been submitted for the award of any other degree of this or any other university.

Gaganpreet Singh

Gaganpreet Singh
801532014
ME (CSE)

This is to certify that the above statement made by the candidate is correct and true to my knowledge.

Kasana

Dr. Singara Singh Kasana
Assistant Professor
Department of CSE
Thapar University
Patiala

Acknowledgement

First of all, I would like to thank the Almighty, who has always guided me to work on the right path of life. This work would not have been possible without the encouragement and able guidance of my supervisor - **Dr. Singara Singh Kasana**. I thank my supervisor for his time, patience, discussions and valuable comments. His enthusiasm and optimism made this experience both rewarding and enjoyable.

I am equally grateful to **Dr. Maninder Singh**, Head of Computer Science and Engineering Department, a nice person, an excellent teacher and a well-credited researcher, who always encouraged me to keep working well and always advised me with his invaluable suggestions. I will be failing in my duty if I do not express my gratitude to **Dr. S.S. Bhatia**, Dean of Academic Affairs, for making provisions of infrastructure such as library, computer labs equipped with Internet facilities, immensely useful for the learners to equip themselves with the latest in the field. I am also thankful to the entire faculty and staff members of Computer Science and Engineering Department for their direct-indirect help, cooperation, love, and affection, which made my stay at Thapar University memorable.

Last but not least, I would like to thank my family whom I dearly love and without whose blessings none of this would have been possible. To my parents, I owe thanks for their care and encouragement. I would also like to thank my close friends for their constant support.

Gaganpreet Singh

(Gaganpreet Singh)

Abstract

A signature is an integral part of a persons individual identity and a mark of his true self. It is the term used to mean the depiction of someone's signing, name or even a single letter which can be any letter or an "A" which is written by a person on a document. While signatures have a widespread acceptance by the public and a huge importance in niche applications like validating documents, validating papers, banking applications and many more makes the signatures an interesting tool of verification. But at the same time it imparts a huge amount of risk in case the signatures of a person are forged so as to access his bank accounts and trespass his property and wealth by the efforts of a skilled forger or intruder. Thus, signature verification has gained recognition from the past three decades so that no frauds could be accomplished by the use of a persons signature. Even with the presence of skilled signature verification algorithms some forgeries are out of scope owing to the skill of the forger or to the change a persons signature witnesses to even a slight extent every time the person signs on a piece of paper. Hence, nowadays signature verification is often combined with the biometric fingerprint verification because two people can produce same signatures but cannot have same fingerprints at all. Still, improvements in the field of signature verification are required and similar kind of work has been proposed in this study where the dataset consisting of both forged and genuine signatures has been used to train various models of machine learning so that the predictability of the fact that which signature is forged and which is genuine is improved to a greater extent. Comparisons have been done on the basis of the situation as to where data preprocessing and dimensionality reduction have been applied and where only the simple data has been used without any changes in the values of the attributes used. The signatures have been converted into vectors on the grounds of the strength of the pixel value in the image of the signature and this study has been applied only on the assessment of off-line signatures. It has shown the maximum efficiency in the terms of accuracy in the case of subspace ensemble of kNN classifier.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

With technological advancements, rapid increments in computing power have taken place. This has led to the ability of computer machines performing complex and computationally intensive algorithms at a faster rate. These developments make automated processes become increasingly popular, targeted potentially at reducing manpower demands. Accurate and rapid programs for matching can thus be written to harness the capabilities of these advancements.

Biometrics in the area of signature matching is not as widely explored as other forms of biometrics. This is natural since mankind has been signing their names as a form of identity verification for thousands of years. Biometric verification has turned into a far reaching methods for anticipation of fraud in money related exchanges and security issues. In particular, handwritten signature verification has been widely used to embrace financial transactions. Due to these requirments signature matching has turned into a prevalent area of research.

1.1 Digital Signature

Signature is the term used to mean the depiction of someone's signing, name or even a single letter which can be any or a "A" is written by a person on any document. Similar to signature there is a term called autograph which is being confused with signature and to clarify, the term autograph is an signature which is artistic. Yet another confusion raise when people have both of them including autograph and signature. As such some

people keep their signatures private while publishing their autograph fully. A behavioral biometric is said to be signature that encodes the ballistic movement of the person which is signing which is a difficult task to imitate.

Where as when compared to the traits which comes in the category of physical traits such as iris, face, finger print etc a higher intra-class and time variability is shown by a signature. While signatures widespread acceptance by the applications in niche and public like validating papers, validating documents, applications of banking etc make the signatures a biometric of interest. Figure 1.1 showing an example of Signature.

Figure 1.1: Signature Example

Signatures can have many meanings and can be used for many purposes like it can be used by a person to indicate the presence of a person physically(e. g. signing in for work), can be used by a person with the intentions to witness(e. g signing of a contract), can be used in approval or authorization as a seal and in the authenticity as a stamp. The way each individual writes and signs is something very personal and often quite distinctive. The main usage of signature verification is to identify the person's identity on the basis of the way he or she signs his or her signature. Originally, signature meant to provide authentication to the document especially in the case of financial transactions such as credit card transactions and bank checks. The person who does signature is a signatory or signer.

1.2 Other Definitions

1.2.1 Histogram

In image analysis to show the distribution of intensities in an image we use a graph called Histogram. To choose an appropriate enhancement operation we can use the information given in the histogram wisely. For example, if the range of intensity values is small as shown by the image histogram so to spread the values across a wider range we can use an intensity adjustment function to do so. A digital image with gray levels in the range [0, L-1]. The histogram of this digital image is a discrete function p(rk) = nk/n, where the count of pixels in the image with that gray level is nk, rk is the k^{th} gray level, the total count of pixels is n and an estimate of the probability of occurrence of gray level rk is given as k = 0, 1, 2, , L-1. p (rk). And a global description of the appearance of the image is provided by the plot of this function for each and every value of k. Figure 1.2 is an example of histogram which is shown below.

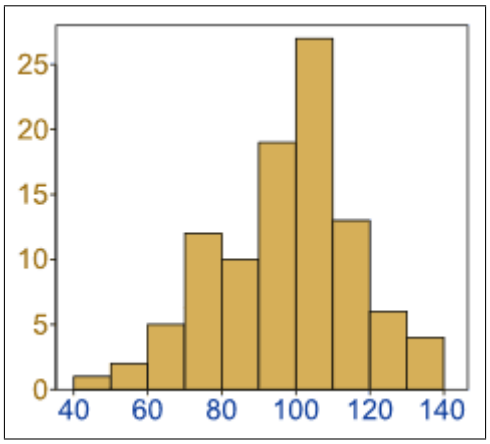

Figure 1.2: Histogram Example

1.2.2 False Acceptance Ratio

The number of forged signatures accepted by the signature verification system with respect to the total number of comparisons made will give us the False Acceptance Ratio.

Calculation of these is show below.

$$\boxed{FAR = \frac{N_{FA}}{N_{FT}} \times 100} \tag{1.1}$$

where N_{FA} and N_{FT} are Number of Forgeries Accepted and Number of Forgeries Tested respectively.

1.2.3 False Rejection Ratio

The number of original signatures rejected by the signature verification system with respect to the total number of comparisons made will give us the False Rejection Ratio.

$$\boxed{FAR = \frac{N_{OR}}{N_{OT}} \times 100} \tag{1.2}$$

where N_{OR} and N_{OT} are Number of Originals Rejected and Number of Originals Tested respectively. A figure 1.3 is the image showing the difference between FRR and FAR.

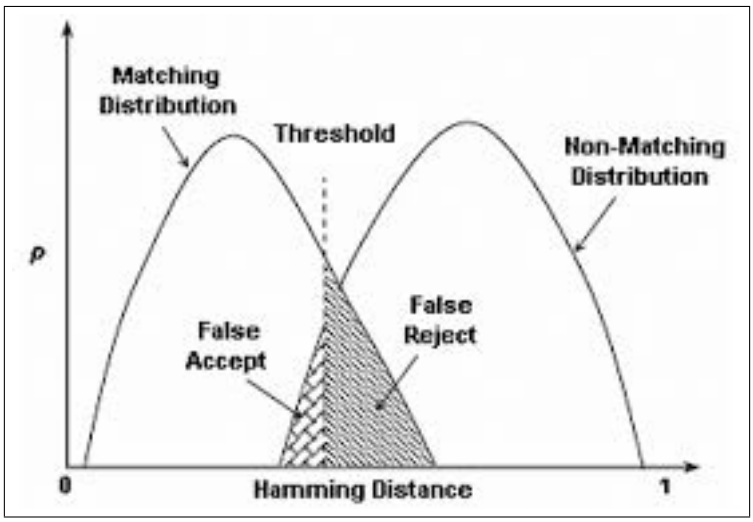

Figure 1.3: False Acceptance and False Rejection Example

1.3 Signature Verification

A necessary task in the society is Signature Verification as for the formal means of personal verification signatures are well established and accepted. Signature verification is a system used by intelligence agencies, banks and high-profile institutions to validate the identity/personality of a person. For other branch capture and to compare signatures in the banks signature verification is oftenly used. For the improvement of the interface between computers and human beings, the recognition of human handwriting is an important concerning and a more economic and attractive man computer interface can be provided if the computer is enough intelligent to understand the human hand writing. In this field a special case that provides us trustworthy means for attestation, authentication, authorization in many high security conditions is signature. The main objective of the verification system for the verification of the signatures is to differentiate the signature between two classes: the original signatures and the forged signatures, which are related to intra-personal and interpersonal variability where Intra Personal Variation is among the signatures of the same person and Inter Personal Variation is between the original signatures and forged signatures. Signature is mostly unreadable and looks like just an image which contains some specific curves and with those curves the writing style of a person is represented that is why verification of signatures is so different with respect to the recognition of characters. A symbol and just a special case of handwriting can said to be a signature so it is necessary and wisdom to just deal with the complete image with special distribution of pixels and the representation of a particular style of writing and not as a collection of words and letters. When the signature saved as image or signature is directly fed into the verification software it is compared to the signature image on file.

The manual verification of the signature just by looking at them is difficult and error prone because of which authenticity can get compromised. Also in case where the data is in large quantity, such signature verification is difficult to implement manually. So automatic handwritten signature verification is required so as to authenticate the user identity. A software can be said as Signature verification and that software compares

signatures and checks for authenticity which can help in saving energy and time and can lead to prevention of human error in the process of signature verification and can bring down the odds of fraud in the process of authentication while verifying signatures. And in the process of verification the verification system computes a confidence score for the signature to be verified and the too low of a confidence score can lead the signature to most likely to be a forged signature.

Although there is a lot of advancement in the technologies since then but still the field of signature verification is the most widely accepted means of authentication when the authenticity of legal documents, bank checks is to be preserved.

1.4 Different Types Of Systems-

There are two systems of signature verification namely **Online signature verification** and **Off-line signature verification**.

Online signature is a biometric used for gaining control of accessing facilities and to get the access of those areas which are protected due to security reasons, personal authentication for security purpose. Figure 1.4 is the example of online signature verification system and the figure depicts the electronic gadget being used for the purpose of capturing the signature. In an online signature verification system, the electronic tablet is used to take the signature data, non-static information about the activity of writing like pressure applied, number of strokes and speed of writing is available while in off-line signature verification systems, signatures which are written on paper as done in the old times are used to be converted to form like electronic form with the help of a scanner or a camera and hence it can be noticed that only the static information is not available.

Figure 1.4: Online System Example

In online verification system each point is related with a corresponding acquisition stamp of time and a co-ordinate of location and other than dynamic compomnents/features for example inclination angles of pen, pressure applied etc that can be captured subject to the equipment used. Online signature verification is the technique which is mainly used for electronic document authentication for various types of applications and also for access control. Due to the differences in the preprocessing, input, classification methods used and feature extraction: online and off-line systems show significant changes in their approaches, especially in the process of preprocessing, matching steps and representation.

Off-line Signature verification is a method of authentication that makes use of the properties like non-static properties of handwritten signature by a person which includes measurement and analyses of the physical activity of signing. Generally, the non-static information includes mainly the style of writing of humans. Since the quantity of information which is available is less, hence the signature verification technique which makes use of off-line techniques is relatively much more difficult. An example image can be shown in the figure 1.5 which is showing the process of capturing the signatures with handwriting as saving the image of signatures using scanning.

Figure 1.5: Off-line System Example

As compared to online signature verification off-line signature verification is more challenging as changes among a user's signatures and easy to examine signature's stance a huge amount of challenge in both off-line and online, information like non static information which is available makes the signature more exclusive and much more challenging to copy in online signature verification. Also, covering both the non-static information and shape of a signature which is online seems to be a bit difficult except for those signatures which are very simple signatures.

Online systems became successful in achieving higher amount of accuracies and which leads the analysts to gain the dynamic knowledge with some success from fixed images. Some special techniques like conoscopic holography, can give the order of stroke and also the applied pressure during signature or handwriting by a pen. These methods are although bulky and hence very high cost equipments and also the approach is also not efficient in time which is difficult to automate. Also these techniques may breakdown with some pen and paper form so such an technique is not practically feasible in the knowledge of making signature verification automatic.
Amongst the above described two types of systems, online and off-line signature verification, this work focusses on off-line signature verification.

Signature authentication cases are also two-fold: while examiners of forensic are more interested in verifying the originality of signer of a document where as many organizations like banks, for routine operations, are more focused in identity control with off-line or

online signatures. In an authentication system of biometrics, the biometric samples are registered by firstly enrolling the users to the system. During the process of verification, a signature with a claimed identity is provided to the system as a query signature and then the signature of the claimed person is taken as reference signature which is then compared with the query signature. The user is authenticated if the computed dissimilarity of those signatures is above a certain threshold, otherwise the user is rejected.

1.5 Types Of Forgeries

There are three different types of forgeries to take into account:

First one is **Random forgery** which is the attempt to copy the signature by the person,where the person don't know the original signature's shape. This person tries to copy the signature only on the basis of his guess and gut considering the letters in the name of the person whose signatures are to be forged.

The second, called **Simple forgery**, which is the attempt by the person to copy the signature who knows the original shape of the signature by don't have much practice of signing.

The last type is **Skilled forgery**, in which a suitable imitation of the original signature model is represented. A different verification approach is required for each type of forgery.

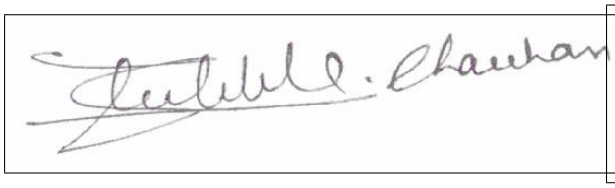

Figure 1.6: Original Signature Example

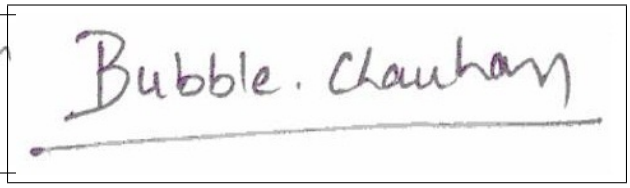

Figure 1.7: Example Of Random Forgery

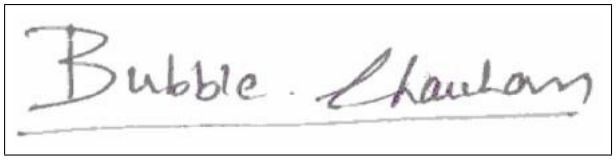

Figure 1.8: Example Of Simple Forgery

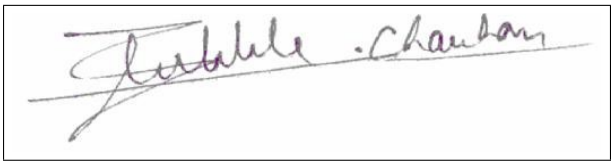

Figure 1.9: Example Of Skilled Forgery

And a another type of signature forgery is determined which is called as **disguised**

signature. In this type, for the purpose of denying ownership in the future, the forged signature is generated by the person himself. For example withdraw of money from an account and then refusing the operation. This category is not yet solved by the researchers as this category poses a difficult problem. However for identifying such forgeries there is forensic interest as well.

1.6 Applications of Off-Line Signature

There are many meanings and purposes of handwritten signatures, as the signatures can be used as signing for the indication of physical presence, signing the contract with the intentions of witness, for the authenticity(e.g stamp), as authorization or as a approval. So there are many applications of off-line signature verification which are as follows.

1.6.1 Financial Institutions

- **Cheques:** In cheques signatures are required as a authentication. As the number of transactions daily are in large number, it is very difficult and it is labour intensive to examine each and every single cheque for the signatures by the bank in detail for the verification of its authenticity and the figure 1.10 is a example figure showing the use of signatures in cheques.

Figure 1.10: figure

Signature's Example On Cheques

This greatly undermines the basic security that consumers expect. Therefore, a

potential remedy for this situation is an accurate off-line signature verification system.

- **Credit Cards:** Off-line signature verification can be used in another area of purchasing the credit card. With the rich use of credit cards, the number of transactions in a day can be very huge, amounting to large amounts of financial transactions based purely on signatures without close audit. The figure 1.11 showing signatures on credit cards. With a fixed signature authentication system, add up in the field of security to the current system can be done through this.

Figure 1.11: figure

Signature's On Credit Card Example

Furthermore, purchasing of credit cards are becoming digitized with just signing on an computerized widget by the customer. Unfortunately, verification of the customer is not done with this widget. All it helps in retrieving the customer's data speedily. However, this gadget can be a stepping-stone for the implementation of a signature verification system since the signatures are captured in the digital form, which makes the identification process more convenient.

1.6.2 Legal Systems

Many legal documents like contracts, land leases and wills, still require signatures as the basic form of authentication. If a system for static signature verification is put in place,

much time can be saved for the verification of the identity of the signer and integrity of the document, thereby enhancing security.

Chapter 2

Literature Review

2.1 Literature

In this chapter, the various researches conducted using the data set consisting of both genuine and forged signatures. From this review it has been tried to find the gaps of the existing techniques, we try to understand the various approaches implemented to derive the best signature verification algorithm. The existing techniques has been classified into online and off-line signature verification techniques.

2.1.1 Off-Line Signature Verification

Porwik *et al.* (2016) proposed a biometric method by using signature's features. Features of a signature are associated with appropriate similarity coefficients and individually matched to a given signature and if necessary, composed features can be reduced. The most promising results are obtained from Hotelling's approach. Data comprising the composed features allow to achieve higher signature recognition level, compared to unprocessed(raw)data. They proposed a new method of data reduction with a new type of similarity measure which gives a high level of signature recognition for various classes of classifiers. They used classifier which is based on Probabilistic Neural Network(PNN) and they also used Particle Swarm Optimization(PSO) procedure to find the optimal parameters of PNN which induced high efficiency. Their signature verification system consists of three units where features are captured, composed features are prepared, data are reduced and verified. The results of proposed system was carried on signatures of the

SVC2004 and MCYT databases and demonstrate the effectiveness in comparison with other methods.

Soleimani *et al.* (2016) presented a method Deep Multitask Metric Learning (DMML) offline signature verification. DMML uses the knowledge from the similarities and dissimilarities between the genuine and forged samples of other classes too. Using the idea of multitask and transfer learning, DMML trains a distance metric for each class together with other classes simultaneously and DMML has a structure with a shared layer acting as a writer-independent approach, that is followed by separated layers which learn writer-dependent factors. They compared the proposed method against SVM, writer-dependent and writer-independent Discriminative Deep Metric Learning method on four off-line signature datasets (UTSig, MCYT-75, GPDSsynthetic, and GPDS960 GraySignatures) using Histogram of Oriented Gradients (HOG) and Discrete Radon Transform (DRT) features. Results of their experiments showed that DMML achieves better performance compared to other methods in verifying genuine signatures, skilled and random forgeries. By employing the idea of multi-task and transfer learning, they mix WI and WD approaches in signature verification to learn a distance metric that measures the similarity between pairs of signatures. This structure helps to use the knowledge form the similarity and dissimilarity of genuine and forged samples of others, to achieve better results. In other words, DMML is a multitask leaning verison of DDML, in which we assume one shared layer as a first layer that is followed by separated layers for all signers. Shared layer helps the network to learn a representation that discovers underlying shared factors among signatures of different individuals, while separated layers try to learn writer- specific factors.

Serdouk *et al.* (2016) proposed a method for off-line signature verification using Artificial Immune Recognition System (AIRS) and it develops representative data that help to discriminate a normal behavior characterized by a given class from abnormal behavior that is characterized by a second class. AIRS is more suitable for applications with detection purposes such as anomaly detection and fault detection than other classifiers like neural networks and SVM. Also, its training algorithm is analytically simple since it is not based on error minimization that is commonly subject to quadratic programming resolution.

Presently, the AIRS is used as the core of the proposed off-line signature verification system, where skilled forgeries are assimilated to antigens that should be detected. Besides, Gradient Local Binary Patterns and Longest Run Features are introduced to respectively, highlight gradient information and pixel distribution in signature images. For generation of features, they used two different descriptors for the generation of signature traits where for the estimation gradient features based on the LBP neighborhood, Gradient Local Binary Patterns are used and in the second descriptor for the description of signatures topology the Longest Run feature is used by considering longest suites of text pixels. For the evaluation of performanceCEDAR and GPDS-100 datasets are used and the results obtained showed that the proposed system has promising performance.

Zois *et al.* (2016) proposed a grid based template matching scheme for off-line signature analysis and verification. This method lies the efficient encoding of the signature's fine geometric structure by grid templates, appropriately partitioned in subsets. Features represented the detection of ordered transitions using lattice shaped probing structures shaped on 5 pixel window binary masks. The verification performance of the method is evaluated on four different signature data-sets producing state of the art results. Additionally, quality characterization of genuine signatures by means of complexity, stability and overall complexity-quality is also carried out. It was observed the both complexity and overall complexity measures correlate strongly with the corresponding opinions expressed by four forensic handwriting experts using the Spear man ranking test. In the proposed work the off-line hand written signature was modeled by focusing to grid based ordered lattices of simple and compound events of pixel assortments. Additionally, a quantifiable way of analyzing signature quality properties by means of complexity, stability and overall quality was also provided by employing the same features which were used for verification. It is expected that modeling of simple and compound grid based events by a sequence of lattice shaped ordered networks will provide sufficient and distinctive fitting to each person's signature model. Validation of the proposed quality measures, particularly complexity and overall quality, is realized by employing a signature database for which four forensic experts have provided their quality opinions. Performance evaluation is carried out on CEDAR and GPDS-100 data sets.

Mustafa Berkay Ylmaz *et al.* (2016) proposed a system based on a score-level fusion of complementary classifiers that use different local features. Each classifier uses a feature-level fusion to represent local features at coarse-to-fine levels. Two different approaches are used for classifiers user-dependent and global classifiers where separately trained classifier for each user is user-dependent classifier and classifier trained with reference signatures of all users and difference vectors of query is called global classifier. The fusion of all classifiers achieves a state-of-the-art performance with 6.97% equal error rate. The GPDS-160 signature database which is publicly available is used for the performance of the experiment. The system based on HOG and LBP features extracted from local grid zones. For either of the representations, features were concatenated in a corse to fine hierarchy which were obtained from grid zones to form the final feature vector. The system was defined as an adapted classification, global fusion and global decision system. It was shown by the system that when enough training data (at least 5 genuine signatures and many random forgeries as the reference set) is available, user-based classifiers are much more successful, then previously observed in the other systems, but user-independent classifiers complement them to improve performance. Results obtained are in par or better as compared to those reported in the past systems for the GPDS database without using any forgeries which are skilled while training.

Ooia *et al.* (2016) proposed a working framework through hybrid methods of probabilistic neural network, principal component analysis and discrete Radon transform which aims to distinguish forgeries from genuine signatures based on the image level. System was experimented on their own independent signature database, and a public signature database MYCT. Equal error rates (EER) of 1.51%, 3.23% and 13.07% are reported, respectively, for random, casual and skilled forgeries of their database. When working on the MYCT signature database, our proposed approach manages to achieve an EER of 9.87% with 10 training samples. Their method gives high accuracy and speed as it extends the previous work by replacing the modeling agent of HMM with PNN where PNN is feasible to distinguish between skilled forgeries and genuine forgeries. A larger database of signatures with forgeries and a powerful specification of PC support were essential to obtain a more reliable statistic of the EER of the system.

Rafal Doroz *et al.* (2016) presented a signature verification method based on the dynamic features of a signature where features space is connected with the set of similarity measures. Features and linked similarity coefficients create a new composed signature features which are then passed to the Hotelling reduction process, where the most appropriate features together with the most distinctive similarity measures are determined for each signature and this process leads to reduction of a composed features space. The approach was checked in different experiments, by considering various classifiers like perceptron and Bayesian network based classifiers as well as k-NN, random trees, random forests and others. In their research they had included the simulation results for the two available datasets of dynamic signatures: SVC2004 and MCYT databases. The results from the research indicate that for data preparation for evaluation of genuine/forged signatures the method processes the input data for various types of classifiers. The method proposes a framework dedicated to verification of signatures, where the data is stored in biometric database. An interesting observation was the fact that processing of data together with Hotelling's reduction method, generates smallest FAR and FRR errors for unreduced input data. The best composed features for a given individual were selected in the Hotelling reduction process and the selected composed features had a biggest impact on the classification phase. These parameters are determined only once for every individual. Correctness and usefulness of thier method had been confirmed by round-robin simulations, where different classifiers have been tested. Efficiency of the classifier was confirmed in the 10-fold cross validation test. Based on reduced data, signature verification method gives also fastest verification time compared to other data structures. The method gives better accuracy and the lower both FAR and FRR errors. Correctness and usefulness algorithm was confirmed by round-robin simulations. The accuracy of the method was very high and was evaluated on the basis of experiments carried out on the wide range of data originating from biometric datasets. The results indicate that in real problems, method of signature verification is valuable and in some cases can out perform the traditional biometric methods, where only raw data are processed.

Cheon *et al.* (2015) proposed a method which uses generalized sparse exponents which uses width-w Non adjacent Forms for signature verification, which can be applied to mod-

ified DSA and ECDSA signatures which further uses tau-adic w-NAF scalars on elliptic curves with complex multiplication such as Koblitz curves. The individual verification factor is accelerated upto 7.49 in the single-signer case and by upto 1.47 in the multiple signer case by the method proposed for 1000 instances over a Koblitz curve K233 and can also be exploited to accelerate batch verification of pairing-based signatures. They significantly improved batch verification of multiple signatures and that improvement can be applied to not only the single-signer case, but also the multiple-signer case. It leads to an interesting research direction to apply their result to various real world applications involving many exponentiations, such as Mix-Net, proof of knowledge, anonymous authentication, and authenticated routing. In addition, it also leads to an interesting issue to develop a fast and provable batch verification method for standard ECDSA signatures which guarantees a sufficient performance gain for single-signer and multiple-signer cases.

Guerbai *et al.* (2015) proposed a system with use of One Class Support Vector Machine(OC-SVM)based on writer-independent parameters, which takes into consideration only genuine signatures and when forgery signatures are lack as counter examples for designing the HSVS which is effective when large samples are available. The OC-SVM is effective when large samples are available for providing an accurate classification. However, available handwritten signature samples are often reduced and therefore the OC-SVM generates an in accurate training and the classification is not well performed. In this system in order to reduce the miss classification modified decision function is used which is done by adjusting carefully the optimal threshold through combining different distances used in OC-SVM kernel. Experiment is conducted on on CEDAR and GPDS hand written signature datasets. The method allows designing the HSVS using few writers and signatures and also defining an only optimal threshold from genuine and fictitious signatures and also allows designing an only optimal threshold from genuine and fictitious signatures derived from the combination of distance used. When a new writer is presented to the system, the same parameters are used without finding the optimal threshold.

Parodi *et al.* (2014) analyzed feature combinations associated with the most commonly

used time functions so as to provide insight on their discriminative power and to quantify the discriminative power a consistency factor is defined. The proposed method is based on Legendre polynomials series expansions for the representation of the time functions as a fixed length associated with the signatures. The expansion coefficients in these series are used as features to model the signatures. The experiment was conducted on publicly available Database which includes two styles of signature namely, Chinese and Western. Two classifiers Random Forests and Support Vector Machines are used in the experiments. The experiments results show a good similarity between the consistency factor computed using only genuine signatures and using skilled forgeries which is a important property as skilled forgeries are not available in the training phase but they do appear when testing the system and also shows a good correlation between the consistency values and the verification errors. It also shows that pen pressure improves the consistency factor and the use of Legendre coefficients as features results in a fixed-length feature vector avoiding the need for length normalization.

Kovari *et al.* (2013) proposed a simplified probabilistic model for off-line signature verification where each verification step can be mathematically described and therefore analyzed and improved. They were able to predict the accuracy of system with a priori known parameters, such as the cardinality and the quality of input samples. With their research they were able to provide answers to several questions, such as why is it so hard to achieve error rates below 10% or how does the number of original samples and features affect the final error rates. The method verified that baseline and loop feature properties could be efficiently approximated with a normal distribution and also shown that these assumptions can be used to define an acceptance threshold for a feature property that minimizes the average error rate. Local features where successfully demonstrated to distinguish genuine and forged signatures. Baselines and loops were found to be good features for characterizing signatures written in Latin writing rather than Chinese writing.

Yu *et al.* (2016) investigated to understand how decision thresholds in the joint decision space of matching and liveness scores in the presence of both spoofing attack and zero-effort attack, with application to signature verification should be optimized which

leads to two dichotomies of methods, namely brute-force approach versus probabilistic approach; and single threshold versus double-threshold approach and further this view leads to three novel methods that have never been reported. A work proposed the use of Artificial Immune Recognition System (AIRS) for off-line signature verification which is an emerging classification method inspired from the learning mechanisms of the natural immune system. Experiments were conducted on public CEDAR and GPDS-100 datasets according to the writer-dependent approach. A parameter independent implementation of AIRS could bbe done by performing simultaneous development of memory cells for all writers. In addition, AIRS performance can be significantly improved if the k-NN classification is substituted by a more reliable decision. And one more practical gap consists of replacing each signature by a vector that contains its dissimilarities with respect to memory cells. Then, dissimilarity vectors of training data will be used to develop a trainable decision function.

Kumar *et al.* (2012) presented a method which includes signature's set of features based on surroundedness property of it and in which the set of feature describes in terms of spatial distribution of black pixels around a candidate pixel, the shape of the signature and the set of features also provides, through the correlation among signature pixels in the neighborhood of that candidate pixel, the measure of texture. The feature set is unique because it contains both shape and texture property. To get a compact set of features various selection techniques of features has also been examined. Two classifiers namely, support vector machine and multi layer perceptron are computed and tested on two available database publicly, namely, CEDAR and GPDS300 corpus signature database. From the results obtained, it was clear that the proposed feature set has got some edge over related features like shape context and auto-correlogram. Performance of the system on both the CEDAR and GPDS300 databases is superior either in terms of accuracy or the time complexity or both, when compared to the state-of-the-art methodologies. Moreover both GPDS300 and CEDAR database is comparable, which indicates that their approach and the feature set are sufficiently general to handle data of varied standards.

2.1.2 Online Signature Verification

Fang *et al.* (2017) proposed a video-based system for in-air signature verification. Despite of having a great dependence on the device of the existing online systems, they proposed a biometrics tends to the long distance and non-contact mode in which, first the fingertip tracking is used for generating unique signature trajectory from the in-air signing. In which combination of improved dynamic time warping method, the Fast Fourier Transform and the analysis on the signature length, a Gaussian distribution-based fusion algorithm is proposed for verification. Finally, comprehensive experiments are conducted on a self-built database consisting of 560 signatures, then a false rejection rate (FRR) of 2.86% and a false acceptance rate (FAR) of 1.90% are achieved with an average matching time of only 24 ms, which have demonstrated the effectiveness of their proposed system. Compared to other algorithms that more than ten training samples are needed, their scheme only requires five genuine samples for training to ensure robust and accurate verification, which is more suitable for the practical signature verification application that less training samples are required.

Sharma *et al.* (2016) presented an enhanced Dynamic Time Warping based online signature verification system by utilizing the code-vectors generated from a Vector-Quantization strategy. They proposed a novel scheme of scoring/voting the aligned pairs in the warping path by a set of code-vectors constructed from a Vector-Quantization step and fuse this score with that of the Dynamic Time Warping, by popular score combination strategies, for verifying a signature. They performed experiments on the publicly available SVC 2004 and MCYT 100 databases. Their procedure gives improved results when compared with other methods for three well acknowledged international signature databases by showing that the probability of correctly classifying a questioned signature is significantly higher when the genuine sample exhibits higher quality. Their work analyzes the distortion values of the warping path of the cost matrix and proposed a score to describe the trend of the distortion values along the warping paths of genuine and forgery signatures also contributed incorporation of the derived score with the normalized DTW score for decision making and inclusion of contextual information to the formulation.

Cpalka *et al.* (2014) proposed a method for the on-line signature verification based on horizontal partitioning where partitions represent areas of high and low speed of signature and high and low pen's pressure where these partitions are more important in the classification process, in which the signatures of the user acquired during training phase are more stable. This experiment is conducted on two datasets namely SVC2004 and BioSecure Database. The method allows to precisely analyse the discriminative power of considered velocity and pressure signals. For individual users weights of created partitions are determined. Values of the weights are proportional to the stability of signing process for individual users. Values of the weights also have a major impact on the classification phase. In this phase more important are partitions which weights of importance have higher value. The method resulted in to analyse the discriminative power, to extract legible knowledge about the process of dynamic signature verification in which knowledge is stored in form of the rules of fuzzy-system. The fuzzy sets appearing in the rules are spaced individually for each user depending on the parameters which describe stability of signing of the user in selected partitions. The rules also take into account the weights of importance of each partition. Correctness of proposed method was confirmed by simulations and the method worked with good accuracy and shows that velocity signal plays more important role than pressure signal in the classification process of dynamic signature. It is therefore a more differentiate signal of individual users and had also shown that the most characteristic for the users are those areas of the signature, which are created during higher value of signing velocity and lower value of the pen pressure.

Signatures recognition systems can be used to identify precisely user identity by making use of signature information such as x, y variations and pressure from a tablet PC. This makes way for using dynamic, i.e. , online handwritten signature based biometric system is more accurate than the static ones, hence can be useful for signature verification applications.

Philip *et al.* (2016) proposed a method with new set of features for online or dynamic signature recognition. Feature vector and their extraction mechanism is implemented using Webber Local Descriptor in this research and this research is further improved

by soft biometric traits of the signature. It includes a cloud based online signature framework. The proposed model was highly scalable, i. e. , during unexpected traffic spikes, the web role and worker role can be scaled up or down to meet demand, while minimizing costs. The proposed online signature system gives 92.50% PI (Performance Index) and 94.25% CCR (Correct Classification Rate).

2.1.3 Hybrid Signature Verification

Radhikaa *et al.* (2015) proposed a signature verification process which focuses on both online and off-line signatures where online signature data is collected by signing and capturing using a web cam where as off-line data is the scanned signatures in form of images. Initially both data undergoes appropriate preprocessing steps and then feature extraction is done which includes pen tip tracking and gradient and projection based features for online and off-line signatures respectively. Online and Off-line methods verifies signatures separately and the results are combined to to be passed to SVM for verification. Both online and off-line approach verifies a signature based on comparison with a previously set threshold value. Finally the results of testing done in these approaches are combined and used by SVM for final verification. The performance of online, off-line and combined approaches have been evaluated and the proposed approach works fairly well.

A varied and unique approach has been implemented by the following researcher where live signatures were taken of people who got drunk after the consumption of alcohol so that the changes can be observed as to how the person signs when he in senses as opposed to how he signs when drunk.

Shin *et al.* (2014) in their research they investigate the detection of alcohol intoxication on the basis of handwritten signatures and evaluate the change in a handwritten signature before and after alcoholic intake. They employed 30 people to evaluate the change in a handwritten signature before and after alcoholic intake. First, they measured the signature verification rate using the online signature verification system and measured using the WACOM Tablet pen before alcohol consumption was 97.0%. Moreover, the size of the characters and the interval between the characters were measured for signatures

collected after alcohol consumption. They detected the level of alcohol intoxication on the basis of the total time taken for writing the signature, the average pressure of the brush, the two-dimensional writing speed, and internal angle of stroke turns. The maximum alcohol detection rate of this method was 95.1%, which was achieved when the examinees were tested 35 min after alcohol consumption. The rate of alcohol detection increases with the alcohol density in an examinees breath. The method is useful because personal authentication system failure is related to the physical condition of examinees, which varies individually. An alcohol detection rate exceeding 90% was possible using the fuzzy logic algorithm.

2.2 Gap analysis

After going through the literature, following gaps have been identified.

1. In one of the studies on-line signature verification uses forward feature selection algorithm to search for the best performing feature subsets with discrete cosine transform which has been applied to 44 time signals, such as position, velocity, pressure and angle of pen. The proposed method shows that The basic on-line features, position signals x(t), y(t) and r(t), are better than other signals in discriminating between genuine signatures and forgeries. Ideally there is a need of features that are stable, i.e. which do not change very much between different genuine signatures, and are hard to forge. For attaining this purpose, capturing signals via a tablet digitizer and using the extracted dynamics information has already been considered in one of the above studies which uses the DCT coefficients in the form of parameters. These features are efficient and experimental results confirm that the proposed method is promising. However, one of the major drawbacks is that humans are not consistent when signing their signatures. Thus the gap lies in focusing on having a lower EER by adding more features that are found useful in other studies which can include the design of a further two stage signature verification system using global and local features.

2. Although the analysis of various works shows that the existing approaches perform

better in the presence of both spoofing and random samples. These approaches may also cause a slight increase in FAR under the conventional zero-effort attack. This calls for a future research direction in better addressing this shortcoming. Apart from this, the current investigation can be extended in a number of ways, which includes the gaps not being limited to the following directions:

- Considering multiple liveness measures.
- Studying the effect of the skill of the forgers in off-line signature authentication case.
- Applying the proposed technique to other biometric modalities.
- Measuring the performance of the proposed framework on several data sets.

Chapter 3

Problem Statement

Signature verification is in dire need of systems which are highly proficient in differentiating between genuine and forged signatures so that no malpractices can be accomplished with the use of forged signatures in the disguise of genuine ones. Proper and careful identification of signatures is required as they are an integral part of a person's identity and are used for the verification of all property-related, legal and banking documents. Rectifying a forged signature as a genuine one can cause frauds and losses of a lots of financial capital duw to which signature verification has been amalgamated with biometric fingerprint verification. There are some gaps in existing signature verification techniques due to which signature verification has to be improved significantly so that not even a slight chance of mistake sustains of declaring a forged signature as a genuine signature. Almost all sectors may it be finance, legal, banking, shopping, parcel delivery, or even the smallest as letter delivery need to use these systems for the accomplishment of high level security. A fully reliable platform of signature verification has to be developed with the minimal error of wrong categorization. The research presented in this dissertation is intended to address the challenge of improving the prediction model to predict the category of the signature which is being examined as the forged or real one and providing timely response in predicting the category. Briefly the important research functions are therefore stated as

- How various data mining techniques can be used in signature verification industry to identify their performance in prediction?
- How does a classification techniques help in developing the prediction model so as to predict accurately the risk of forgery among genuine signatures?

3.1 Objectives of The Research

The basic goal of this research work is to learn the use of varying data mining techniques in the field of signature verification so that the process of signature verification is eased and also improved by the use of this platform. The basic objectives of this research work have been laid as follows:

- To study various classification techniques used in machine learning.
- To propose an improved machine learning framework incorporated with the use of data preprocessing and feature selection to predict if the signature which is been examined is genuine or not.
- To evaluate the performance of proposed framework with the other approaches using the evaluation parameter accuracy.

Chapter 4

Analysis of the Signature Verification Techniques

4.1 Data Mining

Data mining is the process of automatically discovering and deriving useful information in from data which is saved in large data repositories. Data mining techniques help in analyzing large databases so as to find unique and useful patterns that might otherwise not be known to the researchers working in that field. They also provide abilities to predict the outcome also known as the target value of a future observation. Data mining is the integral part of knowledge discovery in databases (KDD), which is the overall process of converting raw data into useful information. The following data mining tasks are associated in this work:

- **Predictive tasks:** The objective of these tasks is to predict the value of a particular attribute (target or dependent variable) based on the values of other attributes (explanatory or independent variables). These tasks refer to the building of a model for the retrieval of the value of the target variable as a function of the explanatory variables being used as attributes. There are two different types of tasks under the category of predictive modelling: classification (for discrete type target variables), and regression (for continuous type target variables).
- **Descriptive tasks:** Here, the main goal of the task is to derive patterns (clusters, correlations, trajectories, anomalies and trends) that collectively or separately recapitulate the fundamental inter-relationships in the data. Since descriptive tasks are often fact-finding by nature, thus techniques of post processing are used to explain

and validate the results that are relevant.

- **Cluster analysis:** This is the technique which is used to find groups of observations that closely related to each other so that the observations that reside in the same cluster are more similar to each other qualitatively than the observations that belong to different clusters.

- **Anomaly detection:** This part helps in identify those observations whose characteristics are considerably different from rest of the data in the dataset. Such observations are known as outliers or anomalies. The basic aim of this procedure is to determine the actual anomalies and avoid the incorrect labelling of the normal objects as anomalous.

4.1.1 Knowledge Discovery and Data Mining

Data mining is an integral part of Knowledge Discovery from Database (KDD) which is the overall process of converting raw data into useful information. The term Knowledge Discovery in Databases, or KDD in short, refers to the elaborated process of extracting knowledge from the raw data, and puts an emphasis on the high level application of particular methods of data mining used for such purpose. It is of great interest to researchers who follow a career in machine learning, databases, pattern recognition, statistics, and knowledge acquisition for expert systems along with the use of artificial intelligence and data visualization. The unifying goal of this process is knowledge extraction from data in the context of extremely large databases which cannot be maneuvered manually. It is done by the use of data mining methods (algorithms) to extract (identify) useful facts and figures in the form of knowledge, using a database along with the required preprocessing, subsampling, and transformations of the data. This process consists of series of steps of transformation, starting from the data preprocessing to post-processing of data mining results so achieved.

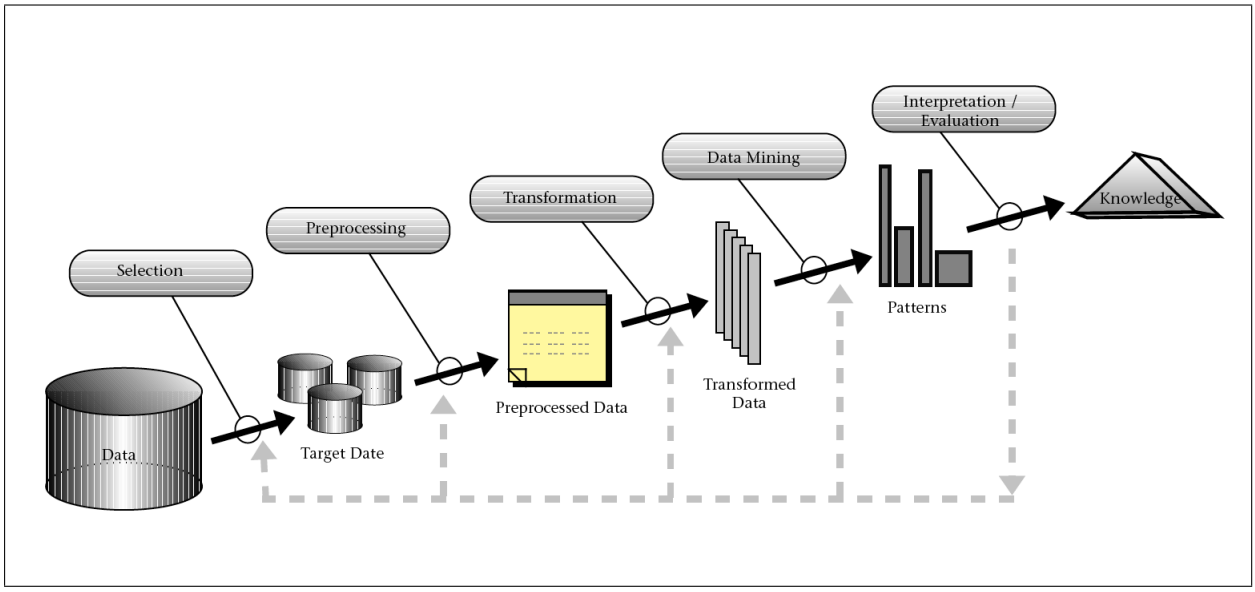

Figure 4.1: Process of Knowledge Discovery from Database

The overall procedure shown in figure 4.1 of discovering and understanding varying patterns from the data involves the iterative execution of the steps as follows:

4.1.1.1 Developing an understanding of

- the domain of the application
- prior knowledge related to the application
- the goals which are to be achieved by the end-user

4.1.1.2 Creating a target data set

Creation of the target data set which refers to the selection of a data set, or focus on the variable subset on which the discovery is to be performed for the data mining results.

4.1.1.3 Data cleaning and preprocessing

- Removal of noise or outliers-

  The scanning hardware and the speckled paper background on which the signature

is signed may introduce certain noises to the signature image. These noises need to be removed so that they will not affect the feature extraction process. In this work, we deploy a median filter to smooth the image, even though we do not compute the real noise distribution. The use of median filtering is similar to an averaging filter like a mean filter. Each output pixel is set to an average of the pixel values in the neighbour hood of the corresponding input pixel. We opt for the median filtering instead of mean filtering due to its more robust average than the mean. More precisely, a single unrepresentative pixel in a neighbourhood will not unduly affect the median value. Due to the fact that median value must be the exact value of one of the pixels in the neighbourhood, it does not create new unrealistic pixel values when the filter straddles an edge. Thus, the median filter is much better at preserving sharp edges than the mean filter.

- Resizing of the signature images to the size of 50×50 -As all scanned signatures are stored in the form of images(as . png files) so we need to make them same as signatures size can vary so as the size of image stored. So in order to get the same size images in order to find the same features from the signatures we used a technique of resizing all the images to some defined size.

- Collecting necessary information to model or account for noise.

- Strategies for handling missing data fields but in this study it was not required.

4.1.1.4 Data reduction and projection-

- Finding useful features to represent the data depending on the goal of the task-The task of feature selection for this particular dataset of signatures is done by the use of HOG features which have explained in detail as follows.
  Histogram of oriented gradients (HOG) is proposed by Dalal and Triggs which includes firstly computing the information at a particular grid zone called gradient

information at each and every pixel where grid zone can be either Polar or Cartesian. And in the second step computation of histogram of gradient orientations in done in that zone. Histogram operates on the thought that it utilize the shape of the signature (coarse shape) by making local directions of gradient with the help of histograms. It is used for the object detection in the field of image processing as a feature descriptor. The HOG method counts the number in a localized zone for the occurrences of gradient orientation of an image. This technique is same as some techniques like scale invariant feature transform, edge orientation and shape contexts but the HOG technique differs in the concept that for uniformly spaced cells it is computed on a dense grid and for the improvement in the accuracy it uses overlapping of local contrast normalization.

It says that distribution of gradient or directions of the edge can be used to depict the appearance of local object within an image and a shape within an image. Cells, which are the small connected zones, are the regions in which the image is divided into. In cell there are pixels and for that pixels histogram of gradient directions is computed. The feature descriptor as told above is the sum up of these histograms and for improving the accuracy intensity across a larger zone of an image is calculated to normalize the local histograms to a block and then using the calculated value to normalize the block by normalizing each cell in the block and with this normalization a better invariance to changes in shadowing and illumination is observed. The HOG algorithm consists of many steps as below:

i Gradient computation

ii Orientation binning

iii Descriptor blocks

- R-HOG blocks
- Circular HOG blocks

iv Block normalization

v Support Vector Machine(SVM)

- The dimensionality of a dataset is said to be the number of attributes that the instances in the data used have. Dataset with a fewer number of dimensions are likely to be different and better than high-dimensional or moderate data in the terms of quality of data. Thus, the complications related to the examination of high-dimensional data are sometimes known as the curse of dimensionality. Thus, a significant inspiration in data preprocessing is dimensionality reduction. When working on the sequential data, it is important to take into account temporal automated correlation; i.e., if two dimensions are close in time, then the values of those dimensions are often very similar and thus one of them can be ignored without affecting the result to any extent. The term dimensionality reduction is often earmarked for those techniques that reduce the dimensionality of a data set by the creation of new attributes that tend to be a combination of the old attributes. The dimensionality reduction by the selection of newer attributes that form subset of the old attribute is also sometimes referred as feature subset selection or feature selection. The advantages of dimensionality reduction are as follows:

    i Mining algorithms perform better if the attributes in the dataset used are lesser.

    ii Dimensionality reduction can lead to the elimination of unrelated features and the noise present.

    iii Can lead to a model which is more understandable and involves lesser number of attributes.

    iv May allow the easy visualization of data.

    v Even if this process is unable to reduce the data to two or three dimensions, data can still be visualized by looking at the pairs or triplets of the attributes so left, and the number of such combinations of attributes is greatly reduced.

    vi The amount of memory and time required by the data mining algorithm is reduced drastically with the reduction in dimensions of the dataset.

  The dimensionality Reduction in this study has been done with the use of **PCA**

algorithm which stands for **Principal Component Analysis.**

It is an algorithm which helps in retaining the most of the variation in the data while reducing the dimensionality of the data where reduction is done by identifying directions in which the variation is maximal in the data. To visualize the similarities and dissimilarities between the samples taken from the data and to see if the samples can be grouped, samples can be plotted.

Principal component analysis (PCA) is a technique for data processing and analysis. It is shown that for a observed data samples the principal axes of the sample data may be computed through maximum likelihood estimation of parameters in a latent variable model.

Listed below are the 6 steps explaining how the PCA operates:

i Dataset containing dd-dimensional samples is taken as whole while ignoring the labels of the class.

ii Calculate the dd-dimensional mean vector.

iii Calculate the scatter matrix of the data set for whole.

iv Calculate eigenvectors and corresponding eigenvalues i.e. $(e_1, e_2, ..., e_d)$ and $(_1, _2, ..., _d)$ respectively.

v Place the eigenvectors by decreasing eigenvalues and choose k eigenvectors with the largest eigenvalues to form a dk dimensional matrix W.

vi Then to transform samples onto new subspace use this dk eigenvector. This can be summarized by the mathematical equation: $y = W_T x$(where x is a d1 -dimensional vector representing one sample, and y is the transformed k1 -dimensional sample in the new subspace.)

4.1.1.5 Choosing the approach for data mining

- To decide if the goal of the KDD process is regression, classification, or clustering, etc.

4.1.1.6 Choosing the data mining algorithm(s)

- Selection of the method(s) which are to be used for the search of patterns in the dataset available.
- Determining which parameters and models may be the most suited here.

4.1.1.7 Data mining

- Search for the patterns which are of high interest in a particular form or representation such as in the form of classification trees or rules, clustering, regression, and so on.

4.1.1.8 Interpreting mined patterns

4.1.1.9 Consolidating discovered knowledge

The data which is to be given as the input can be stored in a number of formats (such as file systems, spreadsheets and relational tables) and they may be kept in a repository which is centralized or be circulated across any number of sites may they be remote or not. The basic aim of preprocessing is the transformation of the raw data into a suitable format for the analysis to follow. The steps involved in data preprocessing have been explained above vividly. Because of the numerous ways available for the data to be collected and stored, data preprocessing appears to be the most arduous task and the most time-consuming step in the overall process of knowledge discovery. Closing the Loop is the phrase mostly used to refer to the practice of integrating the results of data mining into the performance of the decision support systems available. For example, taking the case of business applications, the perceptions made on the basis of the data mining results can be combined with the tools for campaign management so that marketing promotions can be conducted and tested effectively. Such kind of assimilation requires a post-processing step that assures the researcher or the user that only useful and valid results are amalgamated into the decision support system. An important example of post-

processing is visualization which allows the researchers to explore and analyze the data and the data mining results from a variety of viewpoints can be so achieved. Hypothesis testing methods or statistical measures can also be applied during post processing to eliminate false results of data mining.

4.2 Machine Learning

Machine learning is a kind of artificial intelligence (AI) layout that fulfils the technicalities of computers by making them able enough to learn without the computer being explicitly programmed. It is the study of computer algorithms that improve automatically through experience and has been central to AI research since the field's inception. Machine learning is the science of making our computers to act like humans without getting them explicitly programmed. Many researchers and scientists think of machine learning as the best way to make progress on the path of achieving human-level AI. It lays the focus on the creation of computer programs which can improve and change themselves when encountered with new data. Its goals are to learn complicated patterns and to make intelligent decisions based on input data automatically. To deal with a problem in a computer, one first plan an appropriately competent algorithm that deals with the problem and then designs and implements that algorithm in software or hardware. One cannot solve the problem without implement and design an algorithm for that problem. When we are unable to solve a problem manually then Machine Learning extend what can we do with a computer, and how we play with the programmed algorithm.
Machine learning as a logical train inspects the computational premise of learning; consequently, it is basic regardless of the possibility that we are just keen on how people and creatures learn. Machine learning is organized around three primary research foci (Michalski et al. 1997) that are:

- Task-oriented studies The development and analysis of learning systems to improve performance in a pre-determined set of tasks (also known as engineering approach).
- Cognitive simulation The investigation and computer simulation of human learning processes.

- Theoretical analysis The theoretical exploration of the space of possible learning methods and algorithms independent of application domain.

4.2.1 Types of Machine Learning

4.2.1.1 Supervised learning:

This algorithm breaks down the information for preparation and inherently leads to the production of a derived capacity, which can finally be used for the mapping of new cases or instances to be encountered. Supervised learning as regression (for persistent yields) and order (for discrete yields) is an important constituent of Machine Learning. For example, you have input (x) and a yield (Y) and you utilize a calculation to take in the processing capacity called yield with the assistance of this information $Y = f(X)$. The point lies in the exact realization of the registering capacity so well that when you utilize new information (x) you are able to foresee yield factors (Y) for that information. Supervised learning manages to learn a capacity from accessible information. A supervised learning algorithm breaks down the information for preparation and leads to the production of a construed work, which can finally be used for the mapping of new illustrations. Here are some examples are shown below.

- Classification of the e-mails as important or spam
- Labeling of the web pages on the basis of their content
- Voice and Speech recognition.

There are a number of supervised learning algorithms, for example, neural networks (NN), Naive Bayes classifiers and Support Vector Machines (SVMs) etc.
Supervised learning problem can be grouped into classification and regression problem.

- Classification: The goal of the classification algorithm is to predict the target class: yes or no. For predicting two target value or class we use binary classification, i.e. to predict student profile status fail or pass. When we have to predict for more

than two target data class we use the multiple classifications, i.e. considering all the details of the students to estimate which students will earn more points.

- Regression: The goal of regression algorithm is to predict continuous or discrete values. Once in a while, the foreseeing quality can be utilized to locate the straight connection between the attributes. Basic regression algorithm such as linear, polynomial, etc. is used in machine learning problems. Some famous regression algorithm of supervised learning are follows- Linear Regression: It is used to gauge genuine esteems (cost of apartments and houses, the quantity of calls, deals on aggregate etc.) in a way that the opinion of a persistent variable(s) is sustained. In this case, we set up a connection amongst autonomous factors by the fit of the best line found amongst many linear solutions available. This line which fits the best is called the relapse line expressed in a straight form as: $Y = a \times X + b$

4.2.1.2 Unsupervised learning:

The primary aim of this kind of learning is to design and model the basic structure of the data used and to derive knowledge about the distribution of data in orderto learn more about it. Unsupervised learning is a type of machine learning algorithm that draws references from a dataset with input data without labeled responses. It is different from the supervised learning or the reinforcement learning in a way the in this case the learner is trained on the set of data with unlabeled instances. Unsupervised learning technique is further categorized into association and clustering problems:

- Clustering: Clustering is the technique of the arrangement of instances into subsets or sub-populations (also called bunches) so that the instances which are similar to each other or are comparable in some sense belong to the same cluster or group. A clustering issue is a place you need to find the innate groupings in the information, for example, gathering clients by obtaining conduct. It is an approach under the

unsupervised learning technique and is an investigation process for the retrieval of factual information which is to be used in many fields. Following are some real world examples of clustering:

  - In the field of astronomy, with the help of the auto class system a new kind of star was discovered, based upon the clustering of astro-physical measurements.

  - Clustering can be applied in the area of e-commerce where it is common to cluster users into groups on the basis of their web-surfing behavior and purchasing activities. By the clustering results so obtained the merchant can send personalized and customized advertisements to the persons concerned.

- Association: This type of learning is the place where you need to decide the portrayal of expensive segments of your information, for example, individuals that purchase X additionally tend to purchase Y. Some prominent cases of unsupervised learning calculations are:

  - K-means for clustering problems.

  - Association rule mining using Apriori algorithm.

4.2.1.3 Semi-supervised learning:

To defeat the drawback of supervised learning algorithms that they cant make utilization of unlabeled information, semi-supervised learning (SSL) has been proposed to use both marked and unlabeled information. Common approaches to semi-supervised learning include-

- Generative Models

- SVMs

- Graph-Based Algorithms

4.2.2 Various Models

4.2.2.1 Linear Discriminant

Linear Discriminant Analysis(LDA) is the preferable technique for linear classification when we have more than two classes. It is a simple while in application and preparation both and it includes calculated statistical properties for each and every class of our data and it is variance and mean of the variable, if the input is single variable, for each class and the covariance matrix and means are the properties calculated for the multiple variables over the multivariate Gaussian. The estimation of these statistical properties are computed from our data and these estimated values computed are passed to make predictions in the LDA equation and it makes predictions by calculating the probability that the inputs(new set) belongs to each class and the highest probability getting class is the output class and the prediction is made. It uses Bayes Theorem for calculating the probabilities.

$$\boxed{P(Y = x|X = x) = (PIi \times fk(x))/sum(PIj \times fl(x))} \tag{4.1}$$

Where PIi is the base probability observed in the training data of each class (i).

$$\boxed{PIi = ni/n} \tag{4.2}$$

The f(x) above is the calculated probability of x belonging to the class. A Gaussian distribution function is used for f(x) and putting the Gaussian distribution function into the equation shown above and simplifying we came up with equation as shown below and the equation shown below is called a discriminate function and the class is calculated as having the largest value will be the output classification (y):

$$\boxed{Di(x) = x \times (mui/siga^2)(mui^2/(2 \times sigma^2)) + ln(PIi)} \tag{4.3}$$

Di(x) is the discriminate function for class i given input x, the mui, sigma$\hat{2}$ and PIi are all estimated from our data.

4.2.2.2 Logistic Regression

It is a method used for the analyzation of data set where in the data set there are more independent variables than one that determine the output. The output is computed with a variable called dichotomous variable, which have more than two possible results and in this method the variable which is dependent is binary that is it contains 1 (for true, positive, success etc) or 0 (false, negative etc) coded in the data. To fit the best fitting model so that it describes the relationship between a set of variables which are independent and the dichotomous characteristic of interest (dependent variable = response or outcome variable). A formula to compute transformation(logit transformation) of the probability of presence of the characteristic of interest is generated by the Logistic Regression.

$$\boxed{logit(p) = b_0 + b_1X_1 + b_2X2 + b_3X3 + ... + b_KXk} \tag{4.4}$$

where the probability of presence of the characteristic of interest is given by p. The logit transformation is defined as the logged odds:

$$\boxed{odds = \frac{p}{1-p} = \frac{probabilityofpresenceofcharacteristics}{probabilityofabsenceofcharacteristics}} \tag{4.5}$$

and

$$\boxed{logit(p) = \ln(\frac{p}{1-p})} \tag{4.6}$$

It takes some inputs and calculates the probability of some outcome.
For example, if a child has a temperature of 104F (40C) and they have a rash and nausea then the probability that they have chickenpox might be 80%. A rule of thumb in logistic regression, if the probability is ¿50% then the decision is true. So in this case, the determination is made that the child has chickenpox. This is a variation of linear regression, where a model is made to calculate some dependent variable, y, based on some independent variable, x. Then y = mx + b. The model looks for the coefficient m and the y-intercept b. So you end up with some model like the probability of a child having

chickenpox could be something like:

$$\boxed{p(p) = 0.01 \times (temperature) + 0.04 \times (nauseaornot) + 0.03(rashornot)0.4} \quad (4.7)$$

4.2.2.3 K-Nearest Neighbors

It is model with no requirement of learning as it has the ability to store the whole data set. Data can be stored using structures which are complex like k-d trees to make matching and look-up of new patterns more efficient during prediction. Consistency of the training data is to be thought carefully as the whole data set is stored. It might be a good idea to curate it, update it often as new data becomes available and remove erroneous and outlier data and predictions are made directly using the training data set. By searching through the K most similar neighbours and summarizing the result variable for those K samples predictions are made for a new samples and it is mean output for the regression and it is mode in the process of classification. A distance measure is used for determining which is the K instance in the training set more similar to a new input and Euclidean distance is the measure computed for the real-valued input variables. Euclidean distance is calculated as show in the equation 4.8.

$$\boxed{EuclideanDistance(x, xi) = sqrt(sum((xjxij)^2))} \quad (4.8)$$

Other popular distance measures include:

- **Hamming Distance**
- **Manhattan Distance**
- **Minkowski Distance**

4.2.2.4 Decision Trees

Decision tree learning utilizes a decision tree as a predictive model perceptions around a thing (spoken to in the branches) to decisions about the things objective esteem (spoken

to in the clears out). It is one of the predictive modeling approaches utilized as a part of insights, information mining and Machine Learning. Decision Tree algorithm is a very easy technique that is used to make a decision by dividing the inputs into smaller decisions. It is used to predict the target class of the instance of the dataset which is being used just on the basis of much fewer variables given to it as inputs and with the help of a structured decision tree. Like other models, this one also includes mathematics but the mathematics used here is not of a very complex level.

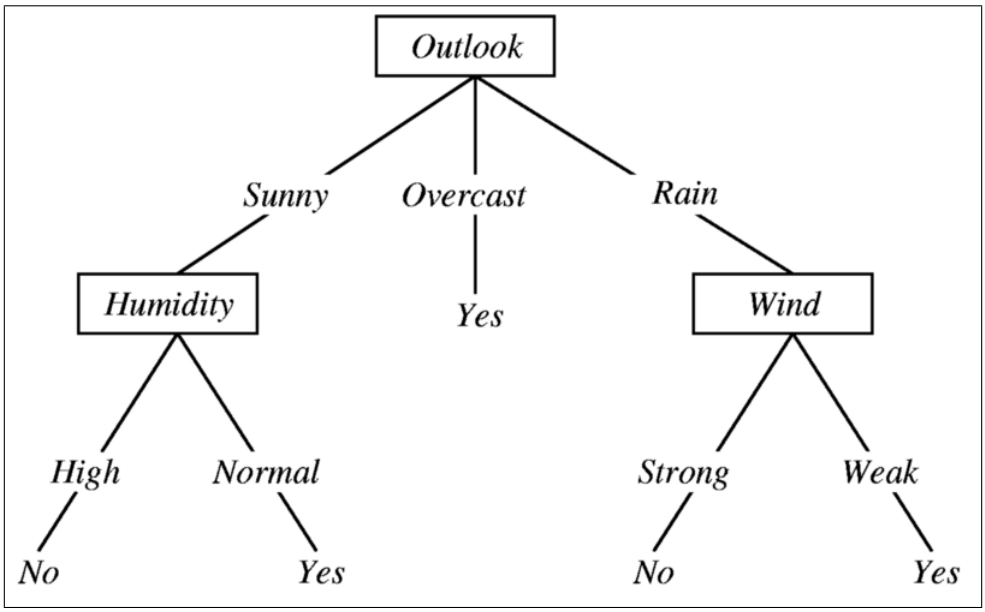

Figure 4.2: Decision Tree Example

Given above is an example of a much simpler decision tree for the purpose of understanding the basic concept in figure 4.2. The goal here is to make a decision on whether the person should play golf on a particular day or not. This is done by making an informed decision on the basis of temperature, wind, humidity and also that the weather is cloudy, sunny or rainy. For example if it is Rainy and the wind is weak then the decision for playing golf will have the outcome yes provided the weather is not humid or mild.
The tree is divided into three kinds of nodes namely internal nodes or the leaf nodes where the condition is the internal node and what decision comes out as the output is the leaf node. The values that the leaf nodes possess are the predefined classes of the dataset. Here in this example the decisions yes or no and the leaves and the factors windy, sunny

etc. are the internal nodes. The root node is the very first nodes and is that attribute from the instance which has been selected as the base to build the decision tree upon. The branches of the tree work as the possible values that the particular internal node or the root node may possess. The most powerful features are the features which get selected earlier during the process of making of the tree and they take the place as the root node or the nodes at the upper levels. The unimportant features either occupy the lower levels or do not find a place in the final decision tree. These trees are built in the top-down fashion and work upon the recursive divide-and- conquer strategy. The making of the tree goes on until a termination criteria is achieved. The process starts by the demarcation of a root node from the input features which has the closest relationship with the output variable. The further nodes are selected by the calculation of the Information Gain (IG) and the formula to calculate this is as follows:

$$\boxed{IG(parent, child) = Entropy(parent) - [p(c1) \times Entropy(c1) + p(c2) \times Entropy(c2)]} \tag{4.9}$$

here $Entropy(cj) = (p(cj) \times log(p(cj)))and p(cj)$is a probability of the child node j. The node having the greatest value of IG is selected as a parent node for the generation to follow. This process is iterated till it gets the leaf node and the tree is completed. There exist a number of algorithms of decision tree which include C4.5, ID3 and CART. Every technique uses a different measure for the selection of the best split so as to find out the most suitable fit to construct the tree.

4.2.2.5 Support Vector Machines

SVM is a model under the category of supervised machine learning which has been explained earlier and is used mainly for the purpose of binary classification. A classification predictor is generated for each test set as input and corresponding output is produced which takes values of the two classes available thus creating a non-probabilistic binary classifier. It is a representation of the illustrations mapped in such a way that the illustrations of distinct categories are bifurcated by a gap which is clear and is as wide as possible. New illustrations are then plotted into the same space and expected to fit

into a category on the basis that on which side of the predesigned gap they fall. This technique hence constructs a linear maximum margin hyperplane in a space with infinite dimensionality, which can further be used for regression, classification and other tasks. It is defined by a weight vector which is denoted by w and bias been represented by b which is the distance of hyperplane from the center. The non-linear separation of dataset is carried out by the use of a kernel function. The classification rule which is used by the SVM classifier is depicted as follows: $Sgn(f(x, w, b))$ and $f(x, w, b) =< w.x > +b$ where f(w,b) presents maximum margin hyperplane for the complex problem and x denotes the example to be classified. Each base classifier which is being used in the generation of the ensemble is trained on the training dataset so as to make them worthy to be used for the prediction of diabetes. The feature space and the predicted class or target labels of the instances of dataset are to the knowledge of each classifier that has been trained, and which ultimately becomes capable of predicting sick and healthy persons from the dataset. The linear margin hyper-plane so found is maximum because a good bifurcation is attained by the hyperplane having the greatest distance to the closest training point belonging to any class (also known as the functional margin), as generally, the greater the margin, the lesser is the error of the classifier.

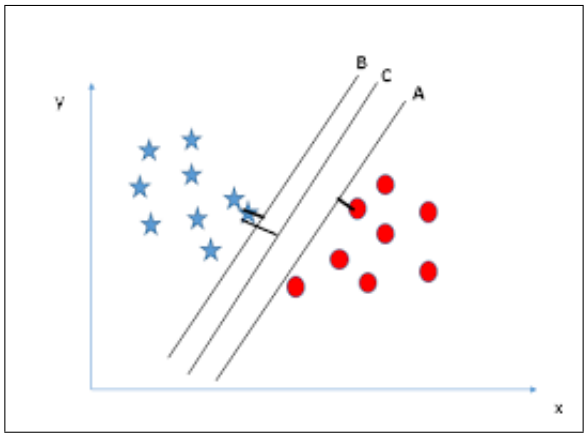

Figure 4.3: Hyperlanes

Lets take the scenario shown in the figure 4.3, where we are dealing with three hyper-planes namely A, B and C and all these hyper planes are partitioning the classes of the dataset very well. Now in order to find out the most appropriate hyper-plane the right hyper-plane right thumb rule is used: Select that hyper-plane as the most suited one which

separates the two classes of the dataset better. And in this case, hyper-plane named B is doing this job perfectly. The primary benefit of SVM is its maximum classification accuracy. It is utilized for pattern recognition and is fundamentally designed for the two-class classification issue. It performs outstanding with the perfect margin for good separation and is really effective in spaces with high dimensionality. It is not good for large datasets because required training time is high. It does not perform well on noisy datasets i.e. target class is overlapping.

4.3 Algorithm

Algorithm 1: Algorithm For The Proposed Method

```
Result: Features File
FolderStack,FileStack,counter,FeaturesFile;
add all folders in FolderStack;
while FolderStack != empty do
    pop one folder from FolderStack and add all the image from that folder to
     FileStack;
    while FileStack !=empty do
        pop one image from the FileStack;
        resize the image to particular standard size as per the proposed method;
        apply median filtering to that image;
        compute HOG features for that image;
        add computed HOG features to FeaturesFile;
    end
end
apply PCA to the FeaturesFile;
run all the models for that FeatureFile for the results;
```

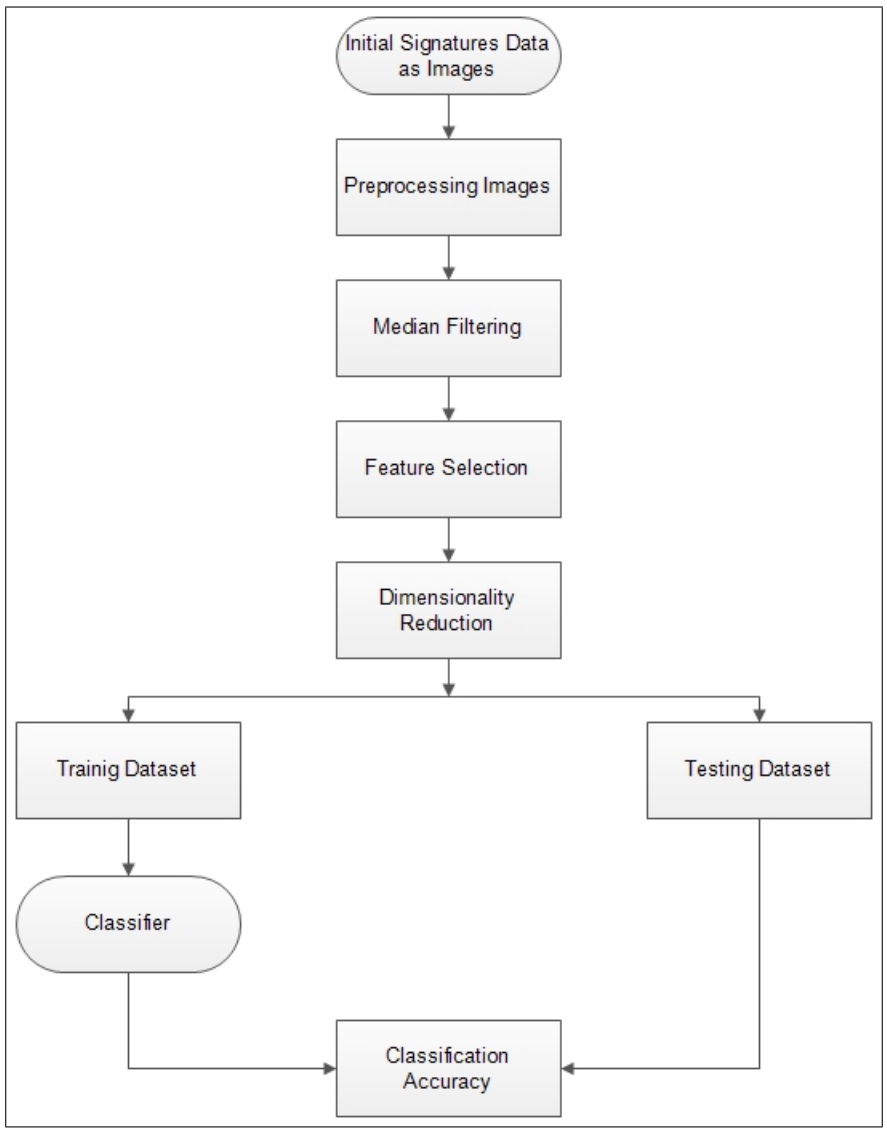

Figure 4.4: Proposed Framework

4.4 Experimental Results and Discussions

For the analysis and the measurement of the performance of the classifiers used, accuracy, is used. They are used because these three measures are more useful in the medical field than any other criteria of measurement. For calculation of accuracy, specificity and sensitivity a confusion matrix is needed which has been explained further. In a confusion matrix: Actual class is the class to which the instance belongs in the original dataset. Predicted class is the class to which the instance is classified by the algorithm used. The

Table 4.1: Table Matrix For The Confusion Matrix

		ACTUAL CLASS	
PREDICTED CLASS		0	1
	0	True Positive(TP)	False Positive(FP)
	1	True Negative(TN)	False Negative(FN)

confusion matrix has been shown in Table 4.1.
TP (True Positive) refers to the number of samples or instances which actually belong to class 0 and also have been correctly classified to class 0 itself. TN (True Negative) refers to the number of samples or instances which actually belong to class 1 and also have been correctly classified to class 1 itself.

FN (False Negative) refers to the number of samples or instances which actually belong to class 0 but have been wrongly classified to class 1. FP (False Positive) refers to the number of samples or instances which actually belong to class 1 but have been wrongly classified to class 0.
In this study, the hybrid approach which includes both median filtering for data preprocessing and PCA for dimensionality reduction for feature selection has shown the maximum accuracy and has proved to be the best in the performance with regard to the prediction of the forged signatures as compared to real ones. The study has made the comparison amongst models which are Decision Tree, Support Vector Machine (SVM), Linear discriminant, Quadratic discriminant, Logic regression, k-nearest neighbor, Boosted ensemble, Subspace discriminant ensemble, Boosted tree ensemble, Bagged tree ensemble, Subspace knn ensemble. These models have been tested in 4 situations which are embarked as follows:

- Ist is the situation which applies neither data preprocessing nor dimensionality Reduction-categorized as SIMPLE.
- IInd is the situation which applies only data preprocessing but no feature selection-categorized as SIMPLE+MEDIAN FILTERING.
- IIIrd is the situation which applies only feature selection but no data preprocessing-categorized as SIMPLE+PCA.

- IVth is the situation which applies both data preprocessing and feature selection- categorized as MEDIAN FILTERING+PCA.

According to the Confusion Matrix stated above, Accuracy is to be calculated as follows: Accuracy= TP+TN / (TP+FP+TN+FN) Accuracy is the measure of how well the classifier is working in predicting the class/target value of the instance in the test dataset as compared to its actual value. Higher the accuracy, better the model is and in this case the hybrid has shown the highest accuracy. Accuracy is actually the weighted arithmetic mean of both Precision and Inverse Precision which are weighted by the Bias present and it can also said to be the weighted arithmetic mean of Recall and Inverse Recall which are weighted by amount of Prevalence present. The experiment is done on the publicly available data base (SigComp2011). The comparison between the proposed method and some other thought methods is shown in the table 4.2.

Table 4.2: Table Showing The Comparison

Model Name	Simply Features	Simple with Median Filtering	Simple with Dimension Reductionality(PCA)	Median Filtering With PCA
Tree(complex Tree)	75. 5	70. 2	69. 9	76. 2
Tree(medium Tree)	74. 8	74. 8	69. 2	76. 2
Tree(simple Tree)	70. 6	72. 0	58. 7	77. 8
Linear Discriminant	83. 1	84. 6	85. 3	89. 5
Quadratic Discriminant	failed	failed	65. 1	66. 4
Logic Regression	62. 2	65. 0	83. 4	86. 0
SVM(linear svm)	88. 1	80. 2	84. 6	87. 4
SVM(fine Guassian)	59. 4	59. 4	70. 6	72. 0
KNN(coarse)	68. 5	74. 1	79. 0	82. 2
KNN(cubic)	62. 3	60. 2	74. 1	76. 2
KNN(weighted)	65. 8	65. 1	75. 9	76. 5
Essembled(boosted)	86. 7	85. 3	78. 9	87. 4
Essembled(subspace discriminant)	83. 2	88. 2	85. 3	89. 5
Essembled(rusboosted Tree)	81. 1	86. 1	76. 2	87. 4
KNN(coasine)	93. 0	93. 5	92. 3	**93. 7**
Essembled(bagged tree)	87. 7	88. 8	81. 1	**92. 5**
Essembled(subspace knn)	95. 1	93. 2	91. 6	**97. 2**

In the table 4.2 the comparison is shown between the different techniques we proposed. Out of the four techniques we proposed the fourth technique comes out with the best results. The different techniques we proposed are as follows:

- In this technique we simply read the image file. Resized the image to the standard size and computed the HOG features for that simple image. With those computed

HOG features we computed the different models and the accuracy of the models in this case is shown in the first column of the table 4.2.

- In this approach we proceeded as the previous approach but with a additional step that is applying the median filtering on the image before computing HOG features for that image and the models on the features with the one additional step of median filtering are computed. The accuracy of the models for this approach are shown in the second column of the table 4.2.

- As the result given by the previous approach were not up to the mark we come up with a new additional step of dimensionality reduction. In this approach we proceeded as the first approach but added one step before computing all the models we applied dimensionality reduction i.e using PCA and the results of this approach are shown in the third column of the table 4.2.

- The results of the previous approach were somewhere good but somewhere not that much good as per all the approaches discussed. So we come up with a mixture of all the previous approaches. We added resizing, median filtering before computing HOG features and added PCA for dimensionality reduction before computing all the models and the accuracy of all the models is shown in the fourth column of the table 4.2.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

This study can be concluded as the introduction of a novel approach which generates an ensemble of both median filtering and dimensionality reduction for the purpose of a platform which conducts the most precise signature verification and gives the best accuracy as and when compared to all the previous approaches used in the same field. It has been shown that, by the use of this hybrid technique, it is possible to predict forgeries and vulnerability in signatures and that too with much reasonable accuracy than the other models used for comparison purposes. The proposed model has rendered high scalability and accountability and it is a far more robust technique than the earlier approaches used for the achievement of same kind of objective may it be in finding the vulnerability of any other kind of forgery or fraud. Classifiers and technique of such type are helpful in the early and correct detection of the onset of frauds in the documents which are at a risk of misuse and a slanderous theft. On the basis of this detection the person concerned can be warned before-hand to change his/her signatures or the detection can be made more fruitful so as to nip the crime in the bud. These improvements will help lower the crime and forgery rates and reduction in the costs of detection as well as classification of the state.

5.2 Future scope

Future scope for the potential research is also discussed in a detailed way that certainly would pave a way for researchers in the future.

- This methodology can be further extended to classify and detect many more types of forgeries such as both online and offline signature complications can be considered in an improved case.
- It can also be used to be extended to an enhanced platform where by the use of an application formed and by making this resource an open-source software every individual can check the authenticity of a signature just by clicking a picture and tallying it with the original signatures available in the database.
- The proposed approach can be improved in its working by the use of ensemble techniques which combine the predictions and classifications of many individual classifiers by the use of bagging, boosting or stacking in a way to aggregate the advantage of so many classifiers into one thereby increasing the accuracy of the system manifolds.

Publications

Gaganpreet Singh and Singara Singh Kasana, *Analysis of the Offline Signature Verification frameworks, The International Conference Recent Advances and Applications in Computer Engineering (RAACE2017).* [Communicated]

Youtube Link

`https://youtu.be/f-4DGOJhnlU`

References

1. Ali J., Khan R., Ahmad N. and Maqsood I., Random Forests and Decision Trees, IJCSI International Journal of Computer Science Issues, 9(5), No 3, 2012.

2. Bhattacharya I. , Ghosh P. and Biswas S. , offline signature verification using pixel matching technique, Elsevier, 2013.

3. Cheon J. H. , Lee M. K. , Improved batch verification of signatures using generalized sparse exponents, Computer Standards & Interfaces, vol.40, pp.42-52, 2015.

4. Cpalka K. , Zalasinski M. and Rutkowsi L. , New method for the online signature verification based on horizontal partionong, Pattern Recognition, vol.47, pp.2652-2661 , 2014.

5. Dietterich and Thomas G, Ensemble methods in Machine Learning, InInternational workshop on multiple classifier systems, Springer Berlin Heidelberg, pp. 1-15, 2000.

6. Doroz R. , Porwik P. and Orczyk T. , Dynamic signature verification method based on association of features with similarity measures, Neurocomputing, vol.171, pp.921-931, 2016.

7. Fahmy Maged M. M, online handwritten signature verification system based on DWT features extraction and neural network classification, Elsevier, 2010

8. Fang Y. , Kang W. , Wu Q. and Tang L. ,A novel video-based system for in-air signature verification, Computers and Electrical engineering, vol.57, pp.1-14, 2017.

9. Hand DJ, Mannila H, Smyth P. Principles of data mining. MIT press; 2001.

10. Karouni Ali, Daya Bassam, Bahlak Samia, Offline signature recognition using neural network approach, , 2010

11. Kovari B. , Charaf H. , A study on the consistency and signature of local features in offline signature verification, Pattern Recognition Letters, vol.34, pp.247-255, 2013.

12. Kumar R. , Sharma J. D and Chanda B. , writer independent offline signature verification using surrondedness feature, Pattern REcogniton Letters, vol.33, pp.301-308,

2012.

13. Nai-Arun N, Sittidech P. Ensemble Learning Model for Diabetes Classification. In Advanced Materials Research 2014 (Vol. 931, pp. 1427-1431). Trans Tech Publications.

14. Parodi M. and Gomez J. C. , Legendre polynomials based feature extraction for online signature verification. Consistency analysis of feature combinations, Pattern Recognition, vol.47, pp.128-140, 2014.

15. Philip J. and Bharadi V. A, signature verification Saas implementation on Microsoft azure cloud, Procedia Computer Science, vol.79, pp.410-418, 2016.

16. Porwik P. , Doroz R. and Orczyk T. , Signatures Verification based on PNN classifier optimized by PSO algorithm, Pattern Recognition, vol.60, pp.998-1014, 2016.

17. Quinlan J. R., Induction of Decision Tree (Reading in Machining Learning, 1986).

18. Radhika K S and S G. , Online and offline signature verification : A combined Approach, Procedia Computer Science, vol.46, pp.1593-1600, 2015.

19. Rashidi S. , Fallah A. and Towhidkhah F. , feature extraction based DCT on dynamic signature verification, Science Iranica, vol.19, pp.1810-1819, 2012.

20. Schapire RE. The boosting approach to machine learning: An overview. In Nonlinear estimation and classification, Springer New York, pp. 149-171, 2003.

21. Serdouk Y. , Hassiba N. and Chibani Y. , new offline handwritten signature verification method based on artificial immune recognition system, Expert Systems With Applications, vol.51, pp.186-194,2016.

22. Sharma A. and Sundaram S. , An enhanced contextual DTW based system for online signature verification using Vector Quantization, Pattern Recognition Letters, vol.84, pp.22-28, 2016.

23. Shin J. and Okuyama T. , Detection of alcohol intoxication via online handwritten signature verification, Pattern Recognition Letters, vol.35, pp.101-104, 2014.

24. SigComp2011 used- `www.iapr-tc11.org/mediawiki/index.php?title=ICDAR_2011_`

Signature_Verification_Competition_(SigComp2011)

25. Soleimani A. , Araabi B. N. and Fouladi K. , Deep multitask metric learning for offline signature verification, Pattern Recognition Letters, vol.80, pp.84-90, 2016.

26. Witten I. H. and Frank E., Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed (USA:Morgan Kaufmann Publishers, 2005).

27. Yilmaz M. B. and Berring Y. , Score level fusion of classifiers in offline signature verification,Information Fusion, vol.32, pp.109-119, 2016.

28. Zois E. N. , Alewijnse L. and Economou G. , offline signature verification and quality characterization using poset oriented grid features,Pattern Recogniton, vol.54, 162-177, 2016.