CHAPTER 1

1. INTRODUCTION

The importance and scope of writer recognition is emerging to gain more prominence in these days. Identification of a writer is exceedingly essential in areas corresponding to forensic expert decision making systems, biometric authentication in information and network security, digital rights administration, document analysis systems and also as a strong tool for physiological identification purposes. The growth of computational engineering, artificial intelligence and pattern recognition fields owes greatly to one of the highly challenged problem of handwriting identification. This thesis titled “Performance Enhancement of Classifiers for Tamil Writer Identification through Modified Support Vector Machine (SVM) Linear Kernel with Parameter Estimation and Deep Learning” presents the implementation of computational astuteness technique to improve discriminative model for writer identification based on Tamil handwritten documents.

1.1. DATA MINING

Data mining is the extraction of hidden predictive information from large databases [1]. It is a powerful new expertise with excessive potential to aid companies focus on the most imperative information in their data warehouses. This implement envisage future trends and activities, permitting businesses to make proactive, knowledge-driven decisions. The automated, suggestive analyses presented by data mining flows beyond the analyses carried out on earlier events rendered by old-fashioned tools, that are distinctive of decision support systems. It is an iterative process within which progress is defined by discovery, through either automatic or manual methods. This technique is most useful in an experimental analysis consequence in which there are no prearranged notions about what will establish an interesting outcome. It is the exploration for original, valuable, and nontrivial information in huge volumes of data. It is a cooperative effort of humans and computers. Superlative results are attained by balancing the knowledge of human experts in describing problems and goals with the search capabilities of computers.

Data mining is one of the quickest growing technologies in the computer industry. Once a trivial interest area within computer science and statistics, it has rapidly expanded into a field of its own. One of the greatest strengths of data mining is reproduced in its wide range of approaches and techniques

that can be applied to a host of problem sets. Since this technique is a regular activity to be accomplished on huge data sets, one of the largest target markets is the entire data warehousing, data-mart, and decision-support community, encompassing professionals from such industries as retail, manufacturing, telecommunications, healthcare, insurance, and transportation. It is concerned with the development and applications of algorithms for detection of a priori unidentified relationships - associations, federations, classifiers from data.

This technique is a process of semi-automatically investigating huge databases to discovery patterns that are effective, original, beneficial, reasonable, also known as Knowledge Discovery in Databases (KDD) [2].

Data mining originates its name from the resemblances among penetrating for cherished business information in a large database. For illustration, finding related products in gigabytes of store scanner data and excavating a mountain for a hint of valuable elements. Equally procedures need either scrutinizing through an massive quantity of material, or cleverly probing it to discovery exactly where the value resides. Particular databases of adequate size and excellence, this technology can produce new business opportunities by providing these capabilities:

- *Automated prediction of trends and behaviors* - Data mining systematizes the procedure of defining predictive information in huge databases. Enquiries that usually required widespread hands-on analysis can now be responded straight from the data quickly. A classic example of a prognostic problem is targeted marketing. This method uses data on past persuasive mailings to recognize the objectives most likely to exploit return on investment in future mailings. Additional applications include forecasting insolvency and other forms of default, and recognizing fragments of a population likely to answer similarly to given events.

- *Automated discovery of previously unknown patterns* - Data mining tools scrutinize through the corpus and in one single step, discover the previously hidden patterns. One instance of pattern discovery includes the evaluation of retail sales data to find the apparently unassociated products, which are frequently bought together. Further pattern discovery obstacles include observing fraudulent credit card transactions and recognizing abnormal data that could characterize data entry keying faults.

As an application, related to additional data analysis applications such as structured queries used in numerous profitable databases or statistical analysis software, data mining symbolize a transformation of kind rather than degree. Several analytical tools exploit a verification-based method, where the user improves a hypothesis and then tests the data to evidence or invalidate the hypothesis. The efficiency of the method can be limited by the originality of the user to improve various hypotheses, as well as the structure of the software being used. In disparity, data mining exploits a discovery approach, in which algorithms can be used to scrutinize numerous multidimensional data relationships simultaneously, identifying those that are unique or frequently represented.

Data Mining Applications

Data mining is finding increasing popularity due to the considerable contribution that can be made by it [3]. It can be utilized for controlling the expenses in addition to contributing towards increase in revenue. Various organizations are consuming this technique which helps to accomplish all phases of the customer life cycle, comprising acquiring new customers, increasing revenue from standing customers and holding worthy customers. By profiling characteristics of good customers, a company can target prospects with related characteristics. By profiling customers who have bought a particular product it can focus consideration on related customers who have not bought that product. By identifying customers who have left, a company can act to maintain customers who are at risk for leaving reducing churn or attrition, because it is usually far less expensive to retain a customer than acquire a new one.

Telecommunications and credit card are two leading companies in applying this method to detect falsified use of their services. Insurance companies and stock exchanges are also involved in applying this technology to reduce fraud. Medical applications are one more fruitful area. This method can be used to envisage the effectiveness of surgical procedures, medical tests or medications. Companies active in the financial arcades use this technique to determine market and industry characteristics as well as to predict individual company and stock performance. Retailers are creating more use of data mining to choose which products to stock in specific stores and even how to place them within a store, as well as to measure the effectiveness of promotions and coupons. Medical firms are mining huge databases of chemical compounds and of genetic material to discover substances that might be useful for development as agents for the treatments of disease.

Data Mining Techniques

As data mining is a normal activity that has to be carried out on massive data sets [4], one among the biggest target markets includes the whole data warehousing, data-mart, and decision-support community, which includes experts from these industries such as retail, manufacturing, telecommunications, healthcare, insurance, and transportation. The most commonly used techniques in data mining are [5] - [8]

- *Rule induction* - Mining of useful if-then rules from corpus based on statistical significance.
- *Classification* - detection of a predictive learning function that classifies records into one of several predefined classes.
- *Regression* - detection of a predictive learning function, which maps a data item to a real-value prediction variable.
- *Clustering* - descriptive task in clustering is to identify a finite set of categories or clusters to label the data.
- *Summarization* - this involves methods for defining a compact explanation for a set or subset of data.
- *Dependency Modeling* - this model describes the significant dependencies between variables or between the values of a feature in a data set or in a part of a data set.
- *Change and Deviation Detection* - learning the most significant variations in the data set.

Trends in Data Mining

As different types of data are available for data mining tasks, so data mining approaches poses many challenging research issues in data mining. The proposal of a standard data mining languages, the growth of effective and efficient data mining methods and systems, the structure of collaborative and cohesive data mining environments, and the applications of data mining to solve enormous applications problems are significant tasks for data mining researches and data mining system and application developers. Here we will discuss some of the trends in data mining that reflect the pursuit of these challenges:

Application Exploration

Earlier data mining was mainly used for business purpose, to overcome the competitors. But in today's scenario data mining becomes more standard in other applications such as biomedicine, stock market, fraud detection, telecommunication and many more. And many new explorations are being done for this purpose. In addition for data mining for business continues to expand as e-commerce and marketing becomes mainstream elements of the retail industry.

Scalable Data Mining Methods

The existing data mining techniques capable of handling only a particular type of data and inadequate amount of data, but as data is expanding at a massive rate, there is a need to develop new data mining techniques which are scalable and can knob different types of data and large volume of data. The data mining methods should be more interactive and user friendly. One important direction towards improving the repair efficiency of the timing process while increasing user interaction is constraint-based mining. This provide user with more control by allowing the specification and use of constraints to guide data mining systems in their search for interesting patterns.

New Methods for Mining Complex Types of Data

The complex types of data like geospatial, multimedia, time series, sequence and text data poses an important research area in field of data mining. There is still a huge gap between the needs for these applications and the available technology.

Web Mining

The World Wide Web (WWW) is enormous collection of globally distributed website contains news, advertisements, consumer records, financial, education, government, e-commerce and many other services. The WWW also contains huge and vibrant collection hyper linked information, providing an enormous source for data mining. Based on the above facts, the web also poses great challenges for efficient resource and knowledge discovery.

Multimedia information is universal and important in numerous applications, and repositories of multimedia are numerous and extremely large. Successively, scholars and expert's requisite new knowledge with upcoming techniques and tools for mining the hidden, useful knowledge embedded

within multimedia collections, thereby helping them discover relationships between the various elements and using this knowledge in decision-making application.

Spatial Data Mining

Spatial data mining is the application of data mining techniques to spatial data [9] [10]. Spatial data mining [11] follows along the same functions in data mining, with the end objective to find patterns in geography. Data mining and Geographic Information Systems (GIS) have its specific approaches, traditions and methodologies to picturing the data. Predominantly, most fashionable GIS have only very basic spatial analysis functionality. The significant explosion in geologically referenced data encouraged by developments in digital mapping, remote sensing, Information Technology (IT) and the global diffusion of GIS emphasizes the significance of developing methodologies to geographical analysis and modeling. Data mining provides automated search for hidden patterns in large databases which can be applied for GIS-based decision-making. In modern times, the mission of combining these two knowledges has develop critical, especially as various public and private sector organizations possessing huge databases with thematic and geographically referenced data begin to understand the huge potential of the information hidden there. Among those organizations are:

- Offices requiring analysis or dissemination of geo-referenced statistical data
- Public health services searching for explanations of disease clusters
- Environmental agencies assessing the impact of changing land-use patterns on climate change
- Geo-marketing companies use spatial location for customer segmentation.

Pattern Mining

Pattern mining is a data mining technique that involves finding existing patterns in data [12]. In this context patterns often means association rules. The benefits identified in association rules motivate to examine supermarket transaction data, that is, to scrutinize customer activities in terms of the purchased products. For example, associations rule "milk powder ⇒ sugar (80%)" states that four out of five customers that bought beer also bought crisps.

Pattern mining is a technique developed as a tool which helps to identify terrorist activity. The National Research Council states pattern mining as Pattern-based data mining looks for patterns that

might be associated with terrorist activity. These patterns act a small signal in a large ocean of noise. Another application like Music Information Retrieval (MIR) plays a major role using pattern mining where patterns seen both in the sequential and non-temporal domains are imported to classical knowledge discovery search techniques.

Subject-based Data Mining

It is also another important data mining technique relating the search for associations between individuals in data. In the environment of fighting terrorism, the National Research Council provides the following definition as: Subject-based data mining customs an scrutiny over individual is measured, depending on the additional information, based on the extraordinary interest, and the objective is to regulate what other persons or financial transactions or movements, etc., are associated to that commencing datum.

Sequence Mining

This technique is concerned with defining patterns based on statistically relevant data as samples where the standards are delivered in a sequence. It is generally assumed that the values are distinct, and thus Time series mining is diligently associated, but generally considered a diverse activity.

There are two different kinds of mining. They are sequence mining, string mining and itemset mining. String mining is extensively used in biology, to examine gene and protein sequences. It is primarily concerned with sequences with a single member at each position. Numbers of prominent algorithms are existing to perform alignment of a query sequence with those existing in databases. The kind of position could also involve corresponding to a query with one subject e.g. matching multiple query sets with each other e.g. Clausal. Itemset mining is used more often in marketing and Customer Relationship Management (CRM) applications, and is concerned with multiple-symbols at each position. Text mining has another popular approach called itemset mining.

Visual Data Mining

It is rightly said a picture is worth a thousand words. So if the result of the mined data can be shown in the visual form it will further enhance the worth of the mined data. Visual data mining is an effective way to discover knowledge from huge amounts of data. The systematic study and development of visual data mining techniques will promote the use for data mining analysis.

There are several key problems within this field is to build corpus and keys for sequence information, mining the frequency of patterns, associating the sequences based on comparison and recovering missing sequence members. The influential apriori algorithm and the more-recent Frequent Pattern (FP) Growth technique are the two common techniques that are applied to sequence databases for frequent item set mining.

1.2. WRITER IDENTIFICATION

Writing is defined as the representation of language in a textual medium through the use of a set of signs or symbols. From the history it describes writing as a significance of political expansion in ancient cultures, which required reliable means for transferring data, retaining financial accounts, observance of historical records and alike. Any language has its history of evolution and development. Languages undergo changes time to time and the recorded thoughts or written form of a language can be an unknown sea if the language becomes extinct or not in use. That is, initially a language is an expression of thoughts by sound, means a spoken language.

On the invention of scripts, written language has been developed, and the evolution goes on. And a distinguishing character is left with each period, rare or tribe. Recognizing or identifying a language of a particular period or of a particular ethnic group has further developed, as language and its purposes grown, to recognize/ identify the writers by the distinctive characteristics of them. Each person has his own manner of writing which depends on a lot of factors like specific shape of letters, spacing between letters, slope, pressure to the paper, average size of letters and so on. Handwriting of a person is also dependent on the mental state of the person like his level of motivation, anger, happiness and others. But it is found that handwriting of a person is relatively stable though may be affected slowly with age.

This uniqueness in handwriting style is exploited in addressing concerns about potential authorship of questioned documents. Recently many crimes have clues in certain inscriptions or handwritten notes. Deciphering the authorship could prove to be the vital turning point in solving/averting danger of such cases. The forensics department considers this branch of study as most

challenging one and many promising research has been done all over the world owing to the fact that there are thousands of scripts in the world.

Most studies about writer identification are based on the documents in English/ Anglo Saxon, Chinese, Arabic, Persian or related languages. With the distinctive characteristics of Indian languages, the tasks on character recognition and writer identification are yet to be developed.

Despite the development of electronic documents and predictions of a paperless world, the importance of handwritten documents has retained its place and the problems of identification and authentication of the writers have been an active area of research over the past few years. Compared to the electronic or printed text, the handwritten text carries additional information about the personality of the person who has written. There exists a certain degree of stability in the writing style of an individual which makes it possible to identify the author for which one has already seen a written text. Writer identification has been used in various applications.

1.2.1. Writer Identification as a Behavioral Biometric

Biometric modalities are classified into two broad categories: physiological biometrics that perform person identification based on measuring a physical property of the human body (e.g. fingerprint, face, iris, retinal blood vessels, hand geometry) and behavioral biometrics that use individual traits of a person's behavior for identification (e.g. voice, gait, keystroke dynamics, signature, handwriting). Writer identification therefore concerns to the kind of behavioral biometrics. From the physical body individual behavior can be observed based on biometric templates are extracted and used in the identification process. Biometric identification is performed by comparing the biometric template measured at the moment when the identification of an unknown person is needed with templates previously enrolled in a database and linked with certainty to known persons.

Physiological biometrics of a person is one of the unique behaviors like fingerprint or iris is strong modalities for individual identification due to the reduced variability and high complexity of the biometric templates used. Hence writer identification comes under behavioral biometrics. Handwritten document analysis is applied in the areas of information retrieval either textually or graphically [13]. However, these physiological modalities are usually more invasive and require cooperating subjects. On the other side, behavioral biometrics is less intrusive, but the achievable enactment is less inspiring

due to the huge variability of the behavior-derived biometric patterns. The identification of a writer based on the handwriting samples still remains a challenging application in forensic field.

1.2.2. Writer Identification in Forensics

Biometric person identification used in forensic labs, automatic writer identification often allows for significant identity in aggregation with the deliberate aspects of a crime, such as in the case of redeem letters. This is a fundamental difference from other biometric methods, where the relation between the evidence material and the details of an offense can be quite remote.

The objective of writer identification systems is less impressive than in the case of DNA or iris-based person identification. In this application, as a rule of thumb, one strives for a near 100% prediction of the correct writer in a hit list of 104 writers, computed from a database in the order of 104 samples, the size of search sets in current European forensic databases. This amount is based on the pragmatic consideration that a number of one hundred suspects are just about manageable in the criminal -investigation process. This target performance still remains an ambitious goal.

The writer detection methods achieved considerable increase in performance and offers strong applicability in forensic field. There exist three groups of script-shape features which are derived from scanned handwritten samples in forensic procedures:

- Completely automated features that are computed from a region of interest present in the image
- Interactively measured features by human experts using a dedicated graphical user-interface tool
- Character-based features which are associated to the allograph detachment which is being produced by each writer.

The whole process of forensic writer identification is not fully instinctive. The features pertaining to groups 2 and 3 require some form of intensive human involvement in executing predefined measuring actions on the script image or in isolating and labeling individual characters or words. Although requiring less human labor, the first group of features has been treated with some skepticism by practitioners within the application domain, given the complexity of the real-life scanned samples of handwriting that are collected in practice.

1.2.3. Writer Identification vs. Handwriting Recognition

Writer identification [14] has its roots within the ancient and extensive automated handwriting identification field. For automated handwriting identification, unchangeable representations are required that have the capability of removing the differences between a variety of handwritings with the aim of classifying the characters' and words' shape with reliability. On the contrary, the issue of writer recognition needs a particular improved representation of these differences that, as such, are distinct to a writer's hand. Owing to its extensive applicability, handwriting identification has always reined the research field involving handwriting analysis.

The objective of handwriting identification is to acquire invariance and generalization. In the case of writer detection, one tries for just the contrary with the objective of maximum exposure to the specificity of the handwriting pattern of every individual for discriminating the writer. It is hugely vital to bring forward the concept that writer recognition could limit few confusions existing in the pattern recognition procedure when information about the common writing habits and distinctiveness of the writer is accessible for the handwriting recognition system [15] [16]. For writer detection, one tries to reveal the specificity of personal handwriting pattern completely for writer differentiation.

Online handwritten [17] digital documents are described as those digital documents, which not just yield information that can be acquired from offline digital documents, but also have temporal information on the handwriting procedure. Such kind of extra information yields substantial cues towards the writers' identity. There must be a clear distinction between writer detection systems and writer verification systems. Writer verification carries out a one on one comparison between a test writer and a database consisting of writers and tries to know about the test writer's authenticity. On the other side, writer detection is involved with the execution of a one on many comparisons and retrieves a ranked list consisting of results obtained after the search. The dissimilarity, even though less, is in the applications where they can be used.

Online document indexing employing writer information yields two unique benefits. First, from the perspective of the information security, writer detection offers innumerable applications in the field of digital rights management and forensic analysis in averting fraud and cases involving identity theft. Secondly, in conditions where massive amounts of documents, forms, notes and minutes of meeting are consistently being processed and dealt with, with the knowledge about the writer's identity would

yield an exceptional value. One among these applications involves the processing and retrieval of the identities of students for the purposes of verification subsequently.

Writer detection mode can be commonly categorized into two kinds- including online and offline. In the case of online, the writing behavior is directly acquired from the writer and transformed into a sequence of signals making use of a transducer device but in the case of offline the handwritten text is utilized for recognition as scanned images. Off-line writer detection is widely regarded to be more challenging compared to on-line as it has additional information regarding the writing style of an individual, including pressure, speed, angle that is not present in the on-line mode.

Off-line writer detection [18] is different from the associated writer verification issue in which one attempts to verify the authenticity of the author who wrote a handwritten sample. There are several applications for offline writer identification in the field of criminal forensics including the evaluation of ransom notes, in which a document of interest is compared with a set of samples from a suspect provided. Various kinds of analysis, recognition, and interpretation can be related with handwriting:

- *Handwriting recognition* - is the capability of a computer to acquire and understand comprehensible handwriting input available as scanned images. It can also be described as the job of changing a language, which is indicated in its spatial form of graphical marks to its symbolic representation.
- *Handwriting interpretation* - involves the task of deciding on the meaning of a body of handwriting, e.g., handwritten address.
- *Handwriting identification* - involves the task of deciding the author of a sample out of a set of writers, supposing that every human's handwriting is distinct. Recognition and verification that find applications in forensic analysis are basically processes, which decide the unique characteristic of the writing of a particular writer, when handwriting identification and interpretation are defined as the processes whose goals are to filter out the changes such that the message can be extracted. The task of reading handwriting is a unique human expertise. Knowledge about the subject domain is necessary as, for instance, in the famous doctor's prescription case, a pharmacist makes use of knowledge about drugs.

1.2.4. Text-Dependent vs. Text - Independent Methods

Writer detection techniques can be classified into two kinds, which include: text-dependent and text-independent techniques. In the case of text-dependent techniques, a writer must write the similar text to carry out identification but in the case of text independent techniques, any text may be utilized for establishing the writers' identity [19].

The text-based techniques are quite identical with signature verification methods and the comparison between each character or word is performed. Therefore, these techniques need the earlier localization and segmentation of the information having relevance. This is generally conducted in an interactive manner with the help of a human.

The text-independent techniques used for writer detection and verification make use of statistical features that are extracted from the complete image pertaining to a text block. A minimal amount of handwriting (e.g. a paragraph with only some text lines) is required for deriving the stable features that are unresponsive to the text content present in the samples. From the application perspective, the significant benefit is that human involvement is reduced. In the case of text-independent methods, the features that are utilized for writer detection yield a wholesome definition about the entire area having the handwriting by removing the location data. It is not desirable to employ text-independent techniques in scenarios where the textual content of the samples is pre-determined and known.

1.2.5. Automatic Writer Identification Framework

Writer detection and verification can be made feasible only when the changes in handwriting pattern between various writers are more than the changes inherent to each individual writer considered separately. Writer identification framework can be categorized into two important methodologies, which include pattern matching and machine learning technique. In the case of pattern matching technique, writer identification system carries out a one-to-many search in a massive database with handwriting samples obtained from known authors and retrieves a possible list of candidates as shown in Fig. 1.1.

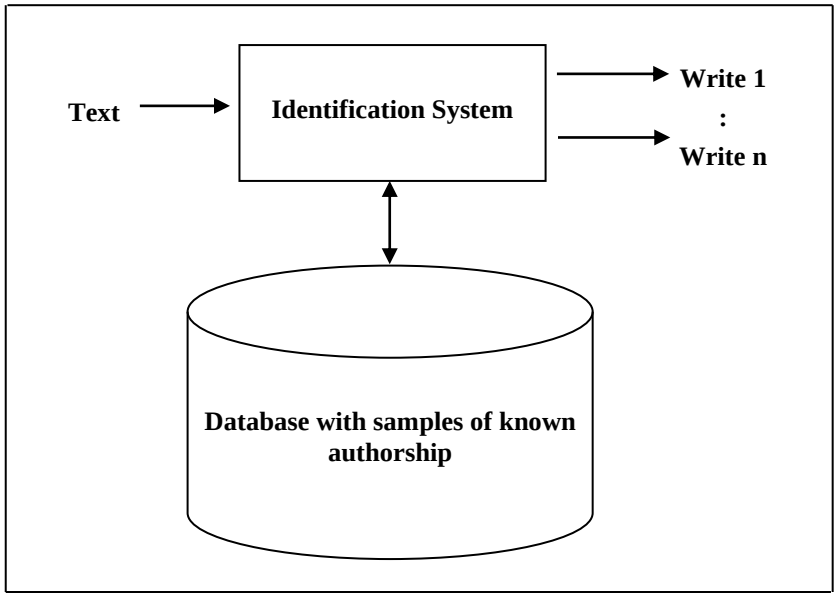

Fig. 1.1 Pattern Matching Approach

Machine learning is a subgroup of artificial intelligence in the computer science field, which frequently makes use of statistical methods to provide computers the capability of learning with data, without any explicit programming. Machine learning is a technique utilized for deriving the sophisticated models and algorithms, which help in prediction that can be categorized into two stages including learning stage and prediction stage. The computer is given example inputs like the writer's handwriting and their suitable Writer Identifier (ID) as training datasets and the learning algorithm will produce a model that can perform the mapping of the new sample handwriting of a writer to the expected result. The generic framework of machine learning based writer detection is illustrated in Fig. 1.2. The available machine learning research work has exhibited its efficacy in writer detection in comparison with pattern matching technique.

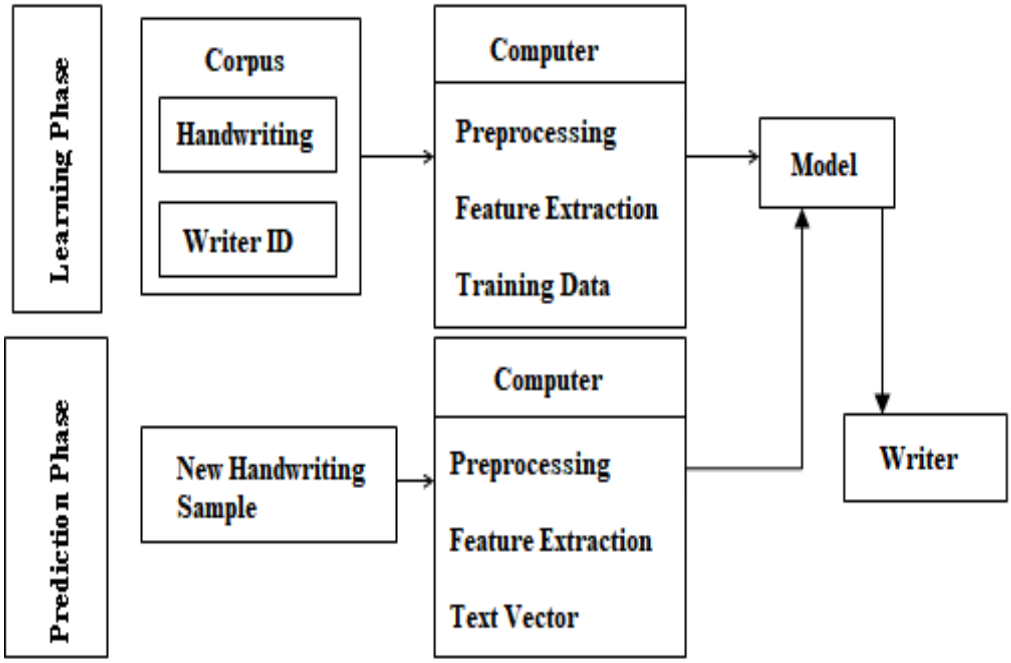

Fig. 1.2 Machine Learning Approach

1.3. NEED FOR AUTOMATIC WRITER IDENTIFICATION

These days, the importance and scope of writer detection is continuing to achieve more prominence. Recognition of a writer is greatly necessary in fields such as forensic expert decision making systems, biometric authentication in information and network security, digital rights management, document analysis systems and also in the form of a potential tool for physiological recognition uses. In the field of forensic science, identification of writer is utilized for the authentication of documents including records, diaries, wills, signatures along with criminal justice. The system of digital rights administration is hugely helpful in protecting the copyrights of electronic media. The writer detection can be primarily applied in the areas discussed below.

***Biometric authentication*:** Writer verification in the form of an authentication system can be utilized for monitoring and regulating the access given to sensitive sites or massive amounts of data, which are consistently being processed and controlled. Knowledge about the writer's identity would be value-addition in information and network security.

- *E-governance and forensic applications* - Writer identification can be used in the form of a tool of support in criminology, historical document analysis and in behavioral biometric authentication in the case of e-governance applications and other relevant IT applications of different private organizations.
- *Signature verification* - Signature is the judicially acknowledged proof of identity of an person in every day operations including automatic banking transaction, legal and commercial environments, electronic fund transfers, document analysis and access control. A signature verification system can identify forgeries and simultaneously, minimize the disapproval of authentic signatures. The handwritten signature introduces several benefits for automated personal recognition.

In general, Writer recognition is vital in forensic and associated branches of science, digital rights management, forensic expert decision-making systems and document analysis techniques for authentication systems and writer verification approaches. The parameters that are usually taken into consideration include universality distinctness, aging, availability, processing complexity and acceptableness. Handwriting samples provide the benefit of being more prevalently available compared to signature samples and also yield more information for analysis of data. It is practically common to make use of handwriting for writer recognition but it is carried out with the assistance of human experts in several cases.

This research work is focused on providing an effective machine learning based solution for the accurate prediction of the identity of a person on the basis of his/her handwriting.

1.4. REVIEW OF LITERATURE

Research works carried out in writer detection has garnered a rekindled interest recently. Different research works have been introduced, which help in distinguishing the writing of different individuals. Depending on the research carried out on a variety of literature works that are available on writer detection, a short report is provided in this section regarding the developments achieved in writer detection in the past few years.

Bulacu and Schomaker (2007) is designed a novel and very efficient methodologies for automated writer detection and verification, which make use of the Probability Distribution Functions (PDFs) acquired from the handwriting images for the characterization of writer

uniqueness [20]. A descriptive characteristic of the techniques was that they were developed to have no dependence on the textual content present in the handwritten samples. The techniques function at two stages of analysis, which include the texture level and the character-shape level. In the texture level, contour-based joint directional PDFs were utilized, which carry out the encoding of the orientation and curvature information to provide an intensive characterization on every style of handwriting. The function associated with these ordinary shapes in a handwriting sample given was unique for the writer and was measured employing a general shape codebook acquired through grapheme clustering. Integrating different features provides an improved performance with regard to writer detection and verification. These newly introduced techniques can be applied to free-style handwriting and they are practically feasible, assuming that some text lines of handwriting are accessible with the aim of getting robust probability estimates.

Bulacu et al (2007) carried out the performance measurement on Arabic handwriting of the text-independent writer identification techniques, which were designed and tested on Western script in last few years [21]. The data from the Institute of Communications Technology (IFN) are used in the experiments and the tests used 350 writers. The results demonstrated that the techniques were quite efficient and the conclusions obtained in earlier studies were valid on Arabic script also. A superior performance was accomplished by integrating textural features and allographic features.

Kannan et al (2008) introduced a novel off-line cursive handwritten recognition system for Tamil that is dependent on Hidden Markov Model (HMM) and a fusion of time domain and frequency domain feature is used [22]. The system's tolerance was obvious since it was capable of surpassing the challenges due to font changes and it was shown to be simple and reliable. A greater level of accuracy in the outcomes was acquired with the realization of this technique on an elaborate database. The results initial were also encouraging and necessitates more research in this area. The outcomes were also promising to investigate the probabilities for following the technique for other Indic scripts also.

Al-Maadeed (2012) introduced a system for text-dependent writer detection on the basis of Arabic handwriting [23]. At first, a database consisting of words was combined and utilized in the form of a test base. Subsequently, features vectors were acquired from the word images of the writers. Before the process of feature extraction, normalization operations were conducted on the word or text line that is being analyzed. In this research work, the author evaluated the feature extraction and

identification operations of Arabic text compared to the identification rate of writers. Since no popular database consisting of Arabic handwritten words were available for researchers for their experiments, a novel database consisting of offline Arabic handwriting text was constructed to be utilized by the research community working on writer identification. The database comprising of Arabic handwritten words gathered from 100 writers was aimed at providing training and testing sets for the research on Arabic writer identification. They assessed the performance of edge-dependent directional probability distributions in the form of features, amongst other features, in Arabic writer detection. The outcomes show that Arabic words and sentences that are longer have a greater effect on writer detection.

Jayanthi and Rajalakshmi technique was introduced to detect a writer from the scanned images pertaining to Tamil handwritten text [24]. The newly introduced strategy was dependent on texture analysis in which handwriting of each writer was considered to be a diverse texture. The technique was text independent and relied on the features acquired from Gray Level Co-occurrence Matrix (GLCM) of the scanned image. The handwriting samples obtained from 70 writers were utilized and the scanning of the documents was done with 150 dpi.

Saranya and Vijaya is designed a model for text-based writer detection over English handwriting [25]. Features were acquired from the scanned images associated with handwritten words and then trained employing Support Vector Machine for the pattern classification algorithm. The SVM with Polynomial kernel achieved an accuracy of 94.27% with writer detection.

Al-Dmour and Abu Zitar (2007) is proposed a novel methodology for feature extraction dependent on hybrid Spectral–Statistical Measures (SSMs) of texture [19]. The outcomes exhibited its efficiency in comparison with multiple-channel filters and the Grey-Level Co-Occurrence Matrix (GLCM) that popular methodologies were rendering a greater performance in writer detection in Roman handwriting. Texture features were acquired for an extensive range of frequency and orientation due to the quality of the distribution of Arabic handwriting in comparison with Roman handwriting, and the most differentiating features were chosen with a model for feature selection employing hybrid Support Vector Machine - genetic algorithm methods. Four classification methods were utilized, which include: Linear Discriminant Classifier (LDC), Support Vector Machine, Weighted Euclidean Distance (WED), and the K Nearest Neighbors (K-NN) classifier. Experiments

were carried out employing Arabic handwriting samples obtained from 20 different persons and quite encouraging outcomes of 90.0% right identification were accomplished.

Karunakara and Mallikarjunaswamy (2011) presented a new technique for the identification of the writer over Kannada language employing Empirical Mode Decomposition (EMD) [26]. They presented a novel technique that tried to recognize the writer assuming that text was constant. Since the handwriting of every writer is visually different from each other; every writer's handwriting shall be considered to be of various textures. These textures were regarded to be a distinct character. These textures were divided employing EMD that in turn creates a sequence of inherent mode functions. The first of the four Intrinsic Mode Functions (IMFs) were taken for research. Therefore, every handwritten image created a four-dimensional vector and was referred to as the Writer features. These features were then stored for identifying the test writer. The k-Nearest Neighbor (k-NN) classifier was utilized for identifying the test writer. The newly introduced technique had been tested over saved features with handwriting features associated with 50 writers inclusive of machine printed ones. Experimental outcomes demonstrated the reliability and flexibility of the newly introduced technique. Using this novel technique, promising experimental outcomes were achieved with the accuracy of Writer 1 - 83%, Writer 2 - 94%, Writer 3 - 90% respectively.

Patil and Shimpi (2011) introduced a novel Character Recognition System, in which generating a Character Matrix and a respective appropriate Network Structure was the key factor [27]. The Feed Forward Neural Network (FFNN) Algorithm provided insight into the inter-operations of a neural network; which is then followed by the Back Propagation Neural Network (BPNN) Algorithm that consists of Training, computing Error, and Changing of Weights. Efforts were made to identify the handwritten English characters by employing a Multilayer Perceptron (MLP) with one hidden layer. Moreover, an evaluation had been performed to decide about the number of hidden nodes to attain a superior performance out of Back Propagation network in the identification of handwritten English characters. The outcomes demonstrated that the MLP networks that are trained by the error Back Propagation algorithm were much commendable with regard to identification accuracy and memory utilization. The outcomes suggested that the BPNN yields a better recognition accuracy of greater than 70% of Handwritten English characters.

Feng and Zhu (2006) carried out the research on the technology of text independent writer detection on the basis of texture analysis on handwritten documents in Chinese [28]. First, during the preprocessing step, the consistent texture images were generated from the input document. A technique for enhanced characters segmentation was introduced relying on the analysis conducted for the character elements and their topological associations. Afterwards, the 32-channel Gabor filter was used for extracting the 64 texture features of writing image by computing the mean values and the standard deviations of filtering the resultant images. At last, multi-class Support Vector Machine classifier was used for meeting out the task of recognition with accuracy 97.91%. The experimental outcome indicates that the approach was efficient and full of potential.

Said et al (1998) tried to evade this presumption by introducing a new algorithm for automated text-independent writer detection from irregularly oriented images of handwriting [29]. Provided that the handwriting of various individuals was frequently visually unique, a global approach is taken up dependent on texture analysis, where the handwriting of each writer was considered to be of a diverse texture. Principally, it let the authors to use any standardized texture recognition algorithm for the task. An accuracy of 96.0% obtained on the classification of 150 test documents obtained from 10 writers was very encouraging. The technique was proven to be reliable to noise and its contents.

He et al (2008) reviewed that the global patterns of the handwritings of a variety of persons were apparently unique and the histogram of the wavelet coefficients of preprocessed handwriting image was capable to be well defined by means of the Generalized Gaussian Model (GGD) [30]. Consequently, in this research work, a novel technique is proposed by integrating the wavelet transform and GGD model for writer detection of handwriting document in Chinese. With the tests carried out by GGD experiment, this technique attained 85.71% accuracy of identification and computational ability as well.

Schlapbach et al (2005) recognized the author of sample handwriting out of a collection of writers, and 100 features were acquired from the sample handwriting [31]. By using feature selection and extraction techniques over this collection of features, subsets of lesser dimensionality were acquired. As observed from the outcomes, considerably better writer recognition rates were achieved when smaller feature subsets that are obtained through diverse feature extraction and selection techniques were exploited. The techniques that taken into consideration in this research work include

Feature Set Search (FSS) algorithms, Genetic Algorithms (GAs), Principal Component Analysis (PCA), and Multiple Discriminant Analysis (MDA). Using 5-NN classifier writer identification rate of 92.08% in the baseline experiment, SBS 94.26% was achieved. Also SFBS and SFFS, produce writer identification rates of 93.17% and 93.44% respectively with SFS of 92.35%.

Dhandra et al (2012) designed a text-independent technique for off-line writer recognition over handwritten documents in Kannada [32]. By making use of every person's handwriting to be a diverse texture image, a group of features dependent on Discrete Cosine Transform (DCT), Gabor filtering and Gray Level Co-Occurrence Matrix, were acquired from the document image segments that were preprocessed. From the implementation it was achieved with the accuracy of DCT with 77%, Gabor with 85.5%, RP1 and GE vector with 84%, Gabor filter response RP2 and GE vector with 88.5%, RP1, RP2 and GE with 82%, GLCM with 9.5%. The results obtained from experiment show that the Gabor energy features offer much more capabilities compared to the DCTs and GLCMs based features for writer recognition from 20 individuals. The experimental outcomes showed the efficacy of the newly introduced techniques and the effectiveness of such global techniques for the writer recognition while being used the document image analysis that, in turn, has much importance in biometrics and forensic science. The short preview of literature survey is given in Table 1.1.

Table 1.1 Summary of Literature Survey

Author	Language	Model	Algorithm	Results
Bulacu and Schomaker (2007)	Arabic Handwriting documents and Western script	Writer Identification and Verification	Feature extraction 1. Textural features 2. Allographic features 3. Writer identification- nearest neighbor classification 4. Writer verification - Neyman Pearson framework of statistical decision theory	99%
Kannan et al (2008)	Tamil script	Off-line cursive handwritten recognition	1. Time domain and frequency domain feature 2. Hidden Markov Model	98% to 92.2% with different lexicon sizes (1K to 20K words).
Al-Maadeed (2012)	Arabic handwriting	text-dependent writer	1. Normalization operations 2. Edge-based directional	73.35%

	documents	identification	probability distributions and other features	
Jayanthi and Rajalakshmi (2011)	Tamil handwritten text	texture independent writer identification system	1. Gray Level Co-occurrence Matrix(GLCM) features	Homogeneity -0.936748 at 135° with distance d =5
Saranya and Vijaya (2013)	English handwritten text	Text-dependent writer identification	Support Vector Machine (SVM)	Polynomial kernel show 94.27%
Al-Dmour and Abu Zitar (2007)	Roman handwriting	Writer identification	Hybrid Spectral–Statistical Measures based feature extraction Classification: Linear Discriminant Classifier, Support Vector Machine, Weighted Euclidean Distance and the K Nearest Neighbors (K-NN) classifier	90.0% correct identification
Karunakara and Mallikarjunaswamy (2011)	Kannada language	Writer identification	Texture analysis - Empirical Mode Decomposition (EMD). Identification - K-NN classifier is used to identify the test writer	Writer 1 - 83%, Writer 2 - 94%, Writer 3 - 90%
Patil and Shimpi (2011)	Handwritten English characters	Character Recognition System	Feed Forward Neural Network Algorithm	Higher than 70%
Feng and Zhu (2006)	Chinese character handwritten documents	Text independent writer identification	1. Gabor filter to extract 64 texture features 2. Multi-class SVM classifier	97.91%
Said et al (1998)	English handwritten documents	Automatic text-independent writer identification	Normalization Feature extraction - Multi-channel Gabor filtering technique and the Grey Scale Co-occurrence Matrix (GSCM) Writer identification - Weighted Euclidean Distance classifier and the	96.0% accuracy

			Nearest Neighbors Classifier (K-NN)	
He et al (2008)	Chinese handwriting document	Writer Identification	Global Wavelet-based features	85.71%
Schlapbach et al (2005)	Different documents	Writer Identification	Feature Selection (FS) - Feature Set Search (FSS) algorithms, Genetic Algorithms, Principal Component Analysis and Multiple Discriminant Analysis	5-NN classifier, we obtain a writer identification rate of 92.08% in the baseline e xperiment. SBS 94.26%. The SFBS and SFFS, prod uce writer identification rat es of 93.17% and 93.44% respectively. SFS - 92.35%.
Dhandra et al (2012)	Kannada handwritten scripts	Text-independent method	Feature extraction - Discrete Cosine Transform, Gabor filtering and Gray Level Co-occurrence Matrix	DCT -77%. Gabor -85.5%, RP1 and GE vector -84%, Gabor filter response RP2 and GE vector - 88.5%, RP1, RP2 and GE - 82%. GLCM - 9.5%.

Motivation

Writer identification is an important issue in many applications like forensic expert decision making, signature verification, biometric authentication, network security and it is highly challenging to produce accurate solution. Currently very less research work has been carried out in Tamil writer identification. Tamil is one of the classical languages having richest literatures in the world. When compared to western scripts and other Indian scripts, Tamil script exhibit a large number of classes, stroke order variation and two-dimensional nature. Hence Tamil language has been chosen for this study of research.

Recent advancements in computational engineering, artificial intelligence and machine learning have proved that it is viable to develop an efficient writer identification model. The models developed based on machine learning techniques are more reliable as they are generated based on intelligent hints collected from the observations. Support Vector Machines have become the method of choice to solve

difficult pattern classification problems. SVM is based on strong mathematical foundations and results in simple yet very powerful algorithms. SVMs find solutions of classification problems that have good generalization performance and able to find linear solutions efficiently using the kernel trick. Deep Learning is an emerging area which takes machine learning into higher level such that the notable attributes or representations are learned automatically with the advent of representation learning. Hence, it is proposed to develop models for writer identification based on Tamil handwriting through machine learning techniques.

In application like signature verification, the identity of an individual is recognized based on single word text i.e., the signature of a person. Whereas the text with more words and paragraphs is required for distinguishing proof of the writer in applications like forensic document analysis. Therefore it is vital to establish the standard model for writer identification. Also increase in the text with more words and sentences will facilitate to capture more distinctive features from the handwritings which are more essential in identifying the writer based on their writing style.

Hence in this work three categories of text such as character, word and paragraph text images have been taken into account and own corpuses have been created. The style and shape of the letters written by the same writer may vary and entirely different for different writers. So the features describing the writing pattern of the writer need to be defined and captured to facilitate the writer recognition task. In this research work global (structural) and local (textural) features of handwritten image have been focused to build accurate models for Tamil writer identification.

1.5. OBJECTIVES OF RESEARCH

Distinctive Handwriting is a thought provoking task in writer identification. Alphabets in the handwritten text may have loops, crossings, junctions, different directions and so on. Therefore exact prediction of individual based on the handwriting is highly complex and challenging task. The main aim of this research work is to propose models for Tamil writer identification to predict the writer accurately through machine learning approach. The major objectives of this research are

- To develop a framework based on supervised learning approach namely Support Vector Machine for Tamil writer identification
- To identify and capture discriminative features from Tamil handwriting and to build SVM classifiers

- To enhance the performance of SVM linear kernel by defining new form of linear kernels using parameter estimation techniques like Weighted Least Square Regression, Bayesian Linear Regression and Principal Component Regression
- To build SVM classifiers with new form of linear kernels such as Weighted Linear Kernel, Bayesian Linear Kernel and Principal Component Kernel to identify writers
- To build Convolutional Neural Network - a Deep Learning framework for Tamil Writer Identification
- To develop a writer identification tool for identifying an individual based on his / her Tamil handwriting

The thesis explains the modeling of writer identification problem as classification task and proposed to build models based on both hand crafted features and self-extracted features through supervised learning and deep learning respectively. These approaches for writer identification exceedingly simplify the traditional writer identification problem and the prediction models are more effective, reliable since it is generated based on intelligent hints collected from the handwriting of individuals.

1.6. ORGANIZATION OF THE THESIS

The rest of the thesis is structured as follows:

Chapter 2 describes about machine learning, supervised classification techniques like Support Vector Machine, deep learning which are adopted in this work for pattern classification.

In chapter 3, the modeling approach used to design writer identification task is addressed. The formulation of writer identification problem as pattern recognition task is explained in detail in this chapter. Also the process of data collection, preprocessing, feature extraction and the framework of writer identification model are described.

The implementation of writer identification models using Support Vector Machine with three standard kernels such as linear, polynomial and RBF are detailed in chapter 4. The performance of the models has been presented with results and findings.

In chapter 5, writer identification models built through modified linear kernel using parameter estimation techniques are elucidated. Various exhibits and findings about writer identification models based on modified linear kernel and their comparison with the performance of the linear kernel are also discussed in this chapter.

Writer identification model built using deep learning approach is described in chapter 6. Convolutional Neural Network have been employed and explained. Results of writer identification using hand crafted features are compared against the results obtained from CNN based writer identification models through self-taught learning and the comparative analysis is presented in this chapter.

In chapter 7, the design and development of the writer identification tool based on the proposed writer identification model using MATLAB GUI is described and the workflow of the tool is elucidated.

Finally, in chapter 8 the research work is concluded by giving an outline of entire research work with various findings. This chapter summarizes the research achievements of the proposed writer identification models and presents recommendations for future research.

CHAPTER 8

8. CONCLUSION

The thesis titled “Performance Enhancement of Classifiers for Tamil Writer Identification through Modified SVM Linear Kernel with Parameter Estimation and Deep Learning” portrays the research work carried on Tamil writer identification through machine learning and deep learning approaches.

The goal of this research is to develop discriminative models for writer identification based on Tamil handwritten text images such as character, word and paragraph using pattern classification techniques. Writer Identification models are built using traditional machine learning approach by identifying and extracting the hand crafted features from Tamil handwritings. Deep learning approach aids in building predictive models through self - extraction of features with improved generalization of writer identification. The proposed approaches considerably make the prediction of individuals more perfect and suggest a beneficial solution by combining local and global features of handwritings.

A set of 100 identical paragraphs written by 300 different writers in Tamil language have been collected as corpus and the handwritten texts are converted into images through scanning. The paragraph text is then segmented into words which in turn segmented into characters. 100 text dependent words are chosen for each writer, creating 30000 word samples. Similarly 100 characters are chosen for each writer, creating 30000 character samples. The images in three different corpuses are preprocessed using various image processing techniques. Feature extraction and feature selection processes have been performed and three independent normalized datasets have been generated.

The foremost task in machine learning approach is aggregating the local and global features from the Tamil handwriting text images and to build writer identification models using Support Vector Machines. In learning Support Vector Machine based classifiers three different kernels such as linear, polynomial and RBF have been employed. The three datasets are partitioned into training and testing sets in the ratio of 80% and 20% and three independent classifiers have been built. Performances of the classifiers are analyzed using various measures such as accuracy, precision, recall, F-measure and time taken to build the model. The results of the experiments shows that the SVM with RBF kernel based writer prediction models are able to attain high prediction accuracy than other kernels.

Linear kernel attains very low accuracy compared to other two kernels. But the observation shows that linear kernel performs faster than the other kernels with less computational complexity. The novelty is introduced in linear kernel and the framework of linear kernel is modified using parameter estimation technique. The proposed weighted linear kernel (WLK) is formulated using weighted least square parameter estimation technique and the writer identification models were built using the same three datasets. The use of WLK-SVM is more beneficial in terms of less computational complexity, minimum time, scalability yielding better results than the linear kernel based SVM method. This work motivated to develop another kernel using Bayesian Linear Regression (BLR).

BLR is used to add the regression parameters to the kernel function in order to reduce error rate. The new form of linear kernel BLK defined using BLR is employed in SVM and the writer identification models were built using BLK-SVM for the same three datasets. The practice of this new form of Bayesian Linear Kernel is more favorable and achieved good performance with less computational complexity.

The next model uses another form of linear kernel formulated based on Principal Component Regression. To define Principal Component Kernel, the principal components of the feature matrix are applied and the coefficients of the linear kernel are determined. The PCK with SVM is trained for three datasets and their performances are evaluated for classification of Tamil writing patterns. The new PCK kernel has comparatively higher performance when compared to linear, WLK, BLK kernels. From the experimental results, it is proved that PCK-SVM achieved better performance with minimum time taken and less computational complexity.

Deep learning is an added significant approach exploited for Tamil writer identification. The subsequent work is intended to adopt Convolutional Neural Networks, a kind of deep learning architecture for Tamil Writer identification. The unified framework of CNN enabled feature learning and classification within the deep learning environment thereby reducing the number of tasks required in traditional learning and to build accurate writer prediction model. Comparison between shallow and deep learning approaches is done by measuring up ANN models against CNN models. It is perceived that the predictive accuracy shown by the CNN models is comparatively higher.

Finally a writer identification tool has been developed by integrating CNN based writer identification model with MATLAB GUI for predicting the identity of writer based on their handwritten Tamil text. Here the tool is developed based on offline text dependent writer identification

which can accept a Tamil handwritten text (either a single word text or a paragraph) as input and can predict the writer of the handwriting. The observations made from this research work are summarized as below.

- Global features extracted from texture of image using Gabor filter and GLCM are not enough to predict the writer as the handwriting contain more structural properties. Hence local features are taken into account to build classifiers
- Global features pooled with local features shows better performance to attains an elevated accuracy in writer identification
- Even though linear kernel achieves very less accuracy compared to other kernels in SVM implementation, it is observed that it performs faster than other kernels and also it shows very less computational complexity
- New form of linear kernels defined using parameter estimation techniques, which enriches the performance of SVM models in predictive accuracy
- Traditional supervised classification technique such as SVM yielded desirable accuracy that motivates to extend the research to next level using deep learning
- In ANN fewer number of hidden units produces high training error and increased number of hidden units shows low training error with reduced generalization ability
- CNN is employed and configured with more number of hidden units which confirms less training error and high generalization power in identification of Tamil writer
- CNN is very effective in absorbing shape variations of handwriting since features of handwritten text images are self-extracted by its architecture
- The research contributions made in this thesis are
- Developed real time corpus as there is no benchmark dataset available in Tamil handwriting
- Identified and captured evocative features from character, word and paragraph handwritten text images

- Built efficient discriminative writer identification models using Support Vector Machine, which has strong mathematical foundation
- Performance enhancement of SVM linear kernel using parameter estimation techniques
- Improved generalization of writer identification through a deep learning architecture- convolutional neural network
- Development of Tamil writer identification tool

It is concluded that machine learning techniques are most suitable in providing solution for writer identification problem. The promising results obtained from this research work have encouraged developing new form of linear kernels for SVM using parameter estimation techniques. Finally, the research is directed to deep learning environment using CNN which is very effective in self-extraction of features and modelling Tamil writer identification. This research recommends the significance in forensic expert decision-making, signature verification and biometric authentication. In future, the work can be extended to develop writer identification model in other Indian languages using text independent handwriting images with emerging computational techniques.