FAKE ACCOUNTS DETECTION IN FACEBOOK USING MACHINE LEARNING TECHNIQUES

Dissertation submitted in fulfilment of the requirements for the Degree of

MASTER OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

PRIYA VIRDI

11509085

Supervisor

Sukhbir Kaur

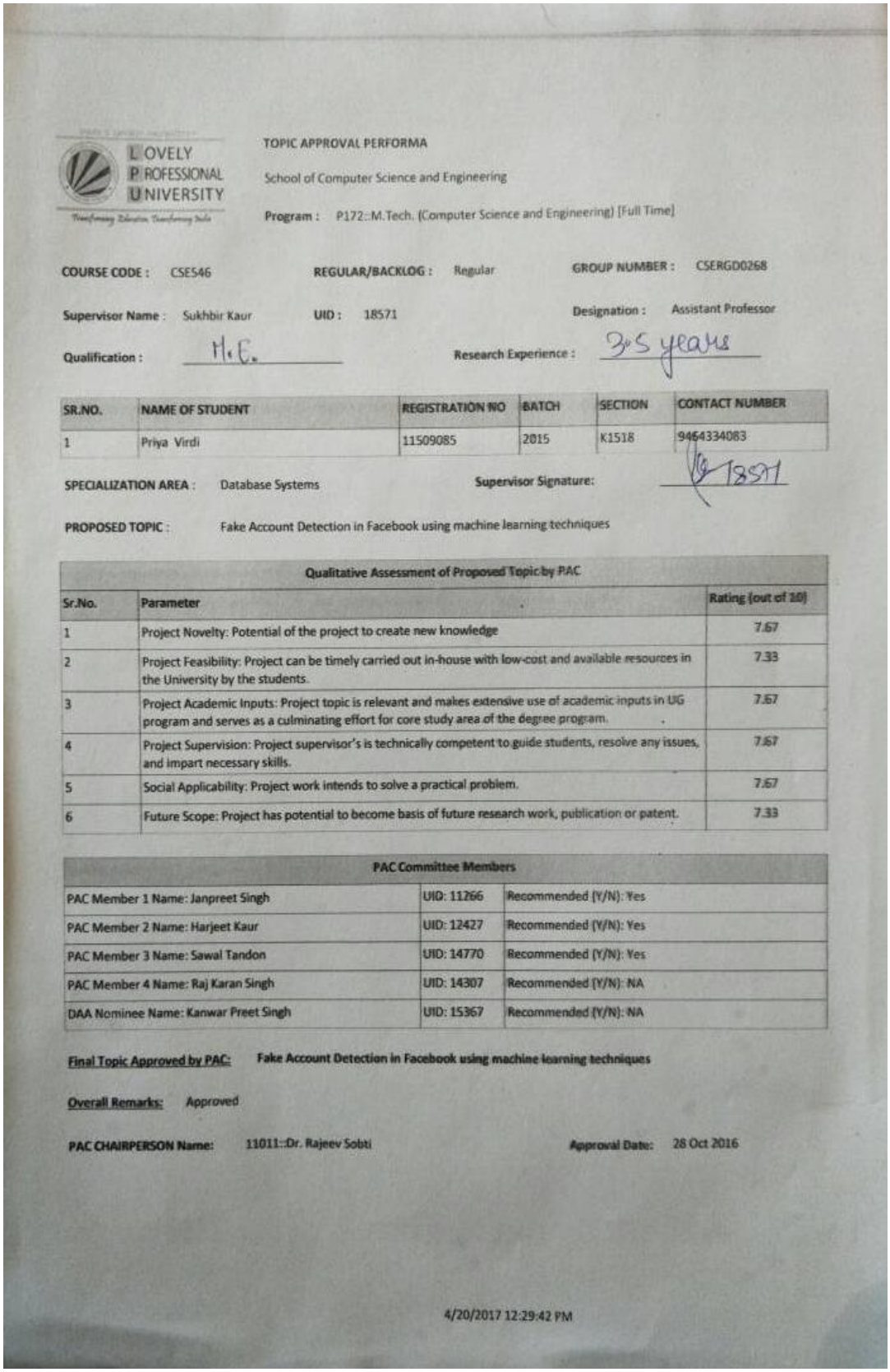
School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

May 2017

PAC Form



LOVELY PROFESSIONAL UNIVERSITY

TOPIC APPROVAL PERFORMA

School of Computer Science and Engineering

Program : P172::M.Tech. (Computer Science and Engineering) [Full Time]

COURSE CODE : CSE546 **REGULAR/BACKLOG :** Regular **GROUP NUMBER :** CSERGD0268

Supervisor Name : Sukhbir Kaur **UID :** 18571 **Designation :** Assistant Professor

Qualification : M.E. **Research Experience :** 3.5 years

SR.NO.	NAME OF STUDENT	REGISTRATION NO	BATCH	SECTION	CONTACT NUMBER
1	Priya Virdi	11509085	2015	K1518	9464334083

SPECIALIZATION AREA : Database Systems **Supervisor Signature:** 18571

PROPOSED TOPIC : Fake Account Detection in Facebook using machine learning techniques

Qualitative Assessment of Proposed Topic by PAC		
Sr.No.	Parameter	Rating (out of 10)
1	Project Novelty: Potential of the project to create new knowledge	7.67
2	Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students.	7.33
3	Project Academic Inputs: Project topic is relevant and makes extensive use of academic inputs in UG program and serves as a culminating effort for core study area of the degree program.	7.67
4	Project Supervision: Project supervisor's is technically competent to guide students, resolve any issues, and impart necessary skills.	7.67
5	Social Applicability: Project work intends to solve a practical problem.	7.67
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	7.33

PAC Committee Members		
PAC Member 1 Name: Janpreet Singh	UID: 11266	Recommended (Y/N): Yes
PAC Member 2 Name: Harjeet Kaur	UID: 12427	Recommended (Y/N): Yes
PAC Member 3 Name: Sawal Tandon	UID: 14770	Recommended (Y/N): Yes
PAC Member 4 Name: Raj Karan Singh	UID: 14307	Recommended (Y/N): NA
DAA Nominee Name: Kanwar Preet Singh	UID: 15367	Recommended (Y/N): NA

Final Topic Approved by PAC: Fake Account Detection in Facebook using machine learning techniques

Overall Remarks: Approved

PAC CHAIRPERSON Name: 11011::Dr. Rajeev Sobti **Approval Date:** 28 Oct 2016

4/20/2017 12:29:42 PM

ABSTRACT

Nowadays social networking sites such as facebook, twitter are incredibly well-liked among people. People interact with their friends via on-line social networks. They share their private and social information using theses social networks. Due to the attractiveness of these websites, a vast number of people uses social networking sites. This fame causes the problem to the websites due to the creation of fake accounts. The owners of fake accounts extract the personal information about other people and spread the forged data on social networks. In our proposed plan, we propose machine learning techniques such as Neural Networks and SVM for detecting the fake accounts on Facebook. Weka tool has been used for the simulation of the algorithm and the obtained results are presented by the proposed plan. Weka is a data mining tool which allows quick user interaction with a simple tool for the identification of fake accounts from provided data. In this, we classify the data using above techniques, which identifies the fake accounts on the social networking sites.

Keywords:- Facebook, Fake Accounts, Feature Selection, Clustering, Classification.

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation entitled "FAKE ACCOUNTS DETECTION IN FACEBOOK USING MACHINE LEARNING TECHNIQUES" in partial fulfilment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor MS. Sukhbir Kaur. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University's Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

Signature of Candidate

Priya Virdi

11509085

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M.Tech Dissertation entitled "**FAKE ACCOUNTS DETECTION IN FACEBOOK USING MACHINE LEARNING TECHNIQUES"**, submitted by **Priya Virdi** at **Lovely Professional University, Phagwara, India** is a bonafide record of her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor
(Sukhbir Kaur)
Date:

Counter Signed by:

1) **Concerned HOD:**

   HoD's Signature: _____

   HoD Name: _____

   Date: _____

2) **Neutral Examiners:**

   External Examiner

   Signature: _____

   Name: _____

   Affiliation: _____

   Date: _____

   Internal Examiner

   Signature: _____

   Name: _____

   Date: _____

ACKNOWLEDGEMENT

I would like to present my deep gratitude to concerned people who helped me out to learn this technology.

I would like to thanks my mentor Ms. Sukhbir Kaur, during this project that helped me as an instructor on each and every step from the day one during the period of Pre- Dissertation. She also helped me in searching and downloading the research papers of good journals related to my proposed plan so I can do my work qualitatively. Her suggestions always help to do my work effectively.

Place : Lovely Professional University Priya Virdi

Date: 11509085

TABLE OF CONTENTS

CONTENTS **PAGE NO.**

LIST OF FIGURE

LIST OF TABLES

CHAPTER 1

INTRODUCTION

On-line social networks are a popular channel to stay in contact. People communicate, and share their everyday activities, photos, and status. Social networking sites like Facebook are very popular among people. To protect the privacy of users on Facebook is a major problem. we propose a technique to protect the privacy of users from fake accounts.

1.1 Introduction

Social networking sites have commonly used the channel of communication between people [1]. Users of social networking sites can share their information and daily activities which attract a number of people towards these sites[2] [3]. One of the most widely used social networking sites is 'Facebook'. Figure 1.1 shows the increasing popularity of Facebook from the year 2004 to 2016. Facebook allow the users to add friends and share various kind of information such as personal, social, political, business etc [3]. Moreover, they can also share photos, videos, travels and another day to day affairs [4]. However, some people don't use these sites with good intent. Therefore they create fake accounts on social networking sites. Fake accounts do not have any real identity. Basically, the person who creates fake accounts is known as Attacker. The attacker uses incorrect information or statistics about some real world person to create a fake account [5]. Using theses fake accounts, attacker spread false information which affects other users [6]. To protect such sensitive data of users is one of the major challenges of social networking sites.

There is a range of machine learning techniques that have been developed to detect fake accounts in social networking sites. Some of these techniques are Neural Network, Naive Bayes, Markov Model and Bayesian Network. In recent researches, it has been found that these techniques make available enhanced results to detect fake accounts. Neural Network consists of many interconnected processing elements. It takes decisions just like a human brain[7]. SVM is supervised machine learning techniques used for classification. It finds the hyper plane to classify the data[8]. Neural network and SVM are able to accept a large amount of random data and suitable to detect the fake accounts

on social networking sites based on various characteristics of accounts. Naive Bayes classifier is based on Bayes' theorem. It predicts the probability that a given variable belongs to particular class [9].

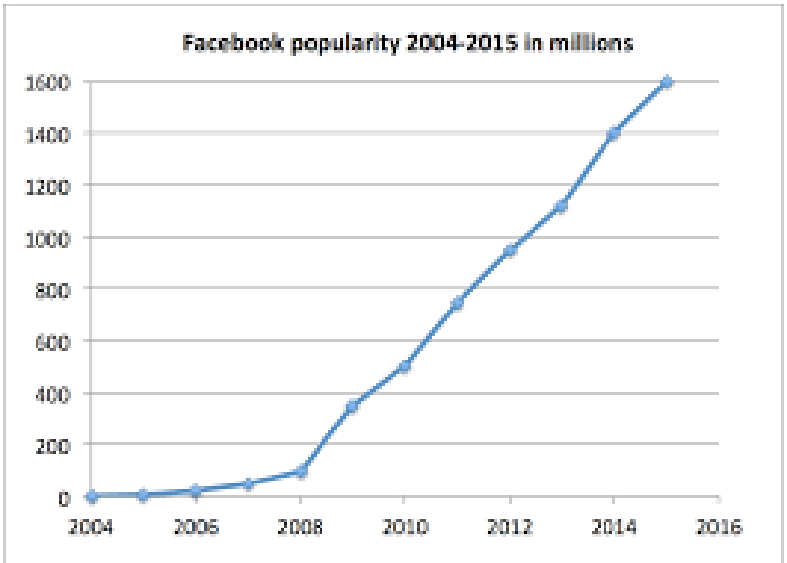


Figure 1.1 Popularity of Facebook from the year 2004 to 2006 in millions

1.2 Classification Techniques

Classification is a data mining technique that allocates objects in a group to target categories or classes. The goal of classification is to perfectly calculate the target class for every case in the data. Classification is two step process. The first step is learning in which classification algorithm analyzed the training data. The second step is a classification in which test data is used to calculate the accuracy of data [9]. Classification predicts the result based on specified input. Item is belonged to which class is calculated by classification algorithms based on the training dataset.[10]. There are various classification techniques are available. Neural Networks and SVM is most successful techniques for classification.

1.2.1 Neural Networks

The basic idea Neural Network is to simulate lots of densely interconnected brain cells inside a computer to make decisions like humans. we don't have to program it to learn explicitly, it learns by itself just like a brain [7]. The basic structure of the human brain is as shown in Figure 1.2.

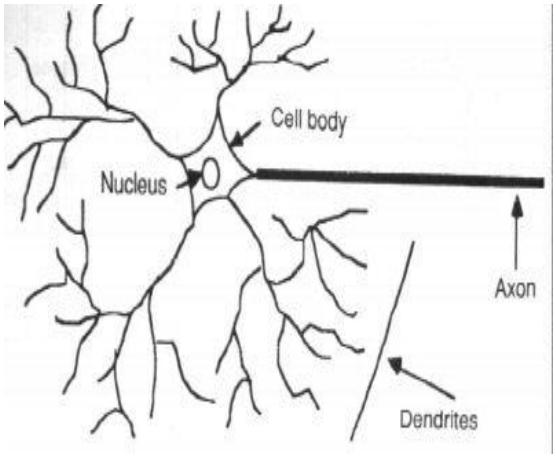


Figure 1.1: Structure of human brain [11]

Human brain contains billions of neurons. Each neuron is made up of cell body with number of connections coming off it, numerous dendrites and single axon. Dendrites are the cell's inputs carrying information towards the cell. Axon is cell's output carrying information away.

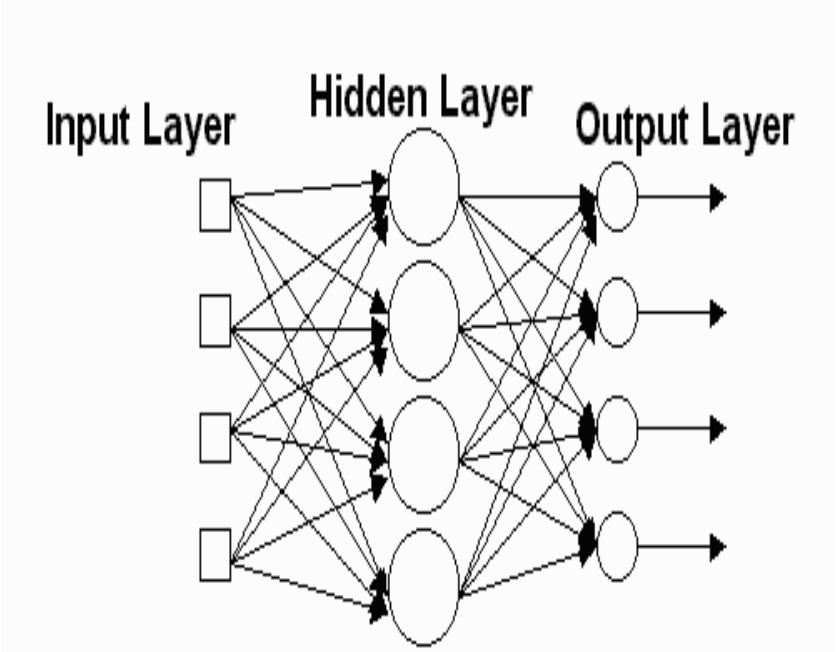


Figure 1.2: Structure of Neural Network

Network has hundreds, thousands, or even millions of artificial neurons called units. Units arranged in a series of layers- Input layer, output layer, hidden layer as shown in Figure 1.3. Input layer accepts the information from the outside, and processing of that

input is done by the hidden layer. The hidden layer plays an important role in producing output. The output layer is used provide the output to the user.

Input layer - receives a various form of information from the outside.
Hidden layer – actual processing is done by hidden layer
Output layer-provides the output.

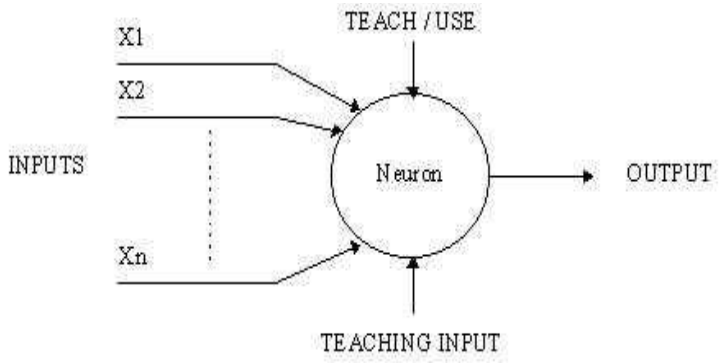


Figure 1.3: A simple neuron

A Neuron has several inputs and one output as shown in Figure 1.4. It works in two different modes- training mode and using mode. In training mode, the neuron is taught to activate for a specific pattern and in using mode, when a given taught pattern is encountered at the input, its affiliate output becomes existent output. Each input is affiliate with a weight. A number multiplied with the input is known as Input weight [12]. The addition is applied to the weighted inputs and if the sum exceeds the predetermined threshold value then neuron fires for the specific input pattern. Mathematically it can be represented as:

$$I1W1+I2W2+I3W3 > T$$

Here I stand for input, W stands for weight and T stands for the threshold.

There are following types of Neural Networks.

Neural Network has two types- Feed forward and back propagation. In feed-forward Neural Network signal travels in a single path, from input to output. It does not give any acknowledgment. However, in back propagation signal travels in both directions to minimize the errors by adjusting weights.

Neural Network also has two learning ways- supervised learning and unsupervised learning. In supervised learning, input and output both are fed. Then the input is processed by the network and matches its actual output with the expected output. The difference between actual output and expected output is called errors. Errors are then propagated reverse by the network, to adjust the weights. In unsupervised learning, the network is provided with only the inputs Figure 1.5 shows the structure of feed forward and feedback network. The system must then arrive at conclusion independently which characteristics could be used to cluster the input data [7][13].

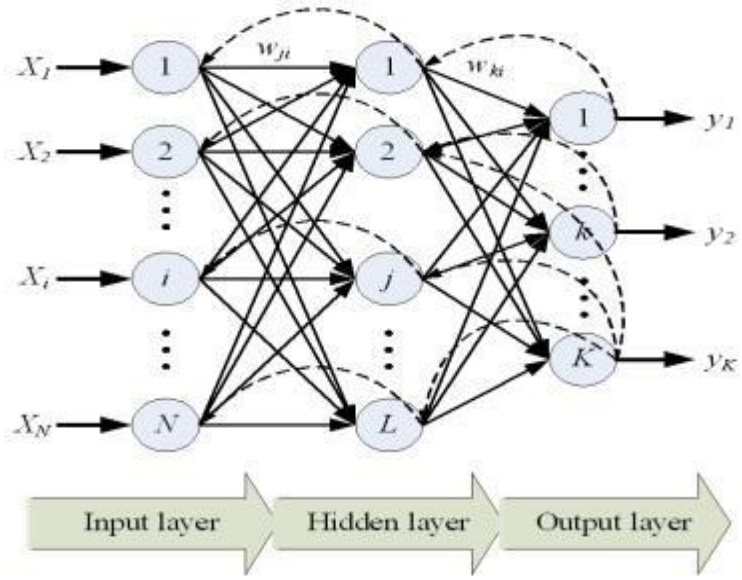


Figure 1.4: Feed forward and Feedback network [14]

1.2.2 Support Vector Machine (SVM)

SVM (support vector machine) is supervised machine learning technique used to classify the data. SVM provides higher accuracy than the other classification techniques. The basic idea behind SVM is to maximize the margin of data by discovering the best possible separating hyper plane. In the fig 1.3 there are multiple hyper planes which separate the data but best one is chosen [8]. After that to get the margin, the difference between hyper plane and the closest data point is computed and double this value. There are multiple hyper planes in Figure 1.6. But choose the best hyper plane which helps to classify the data accurately. SVM provides higher accuracy than the other classification techniques.

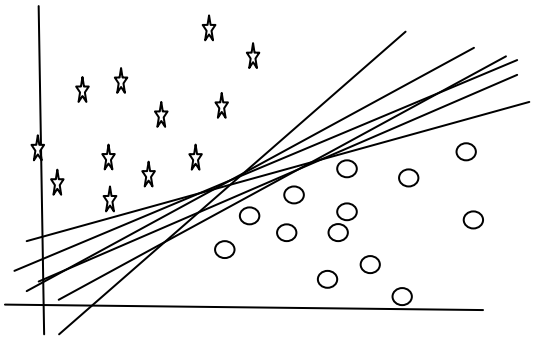


Figure 1.5: Separating hyper planes

Given a particular hyper plane, we can compute the distance between the hyper plane and closest data point. Once we have this value, if we double it we will get margin. There will never be any data point inside the margin as shown in Figure 1.7.

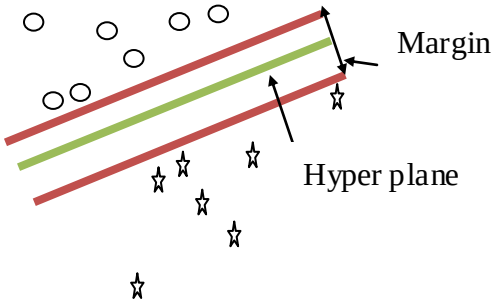


Figure 1.6: Optimal hyper plane and margin

There are basically two types of SVM as follows:-

I. Linear SVM: - In linear SVM the plane is linear. Figure 1.8 shows the linear plane, In which there is one hyper plane separating the data. One side of data belongs to one class and other side data belongs to another class. In linear SVM data is simply separable. There is no complexity while the separation of data into classes like nonlinear SVM. Therefore it used only for simple classification of data.

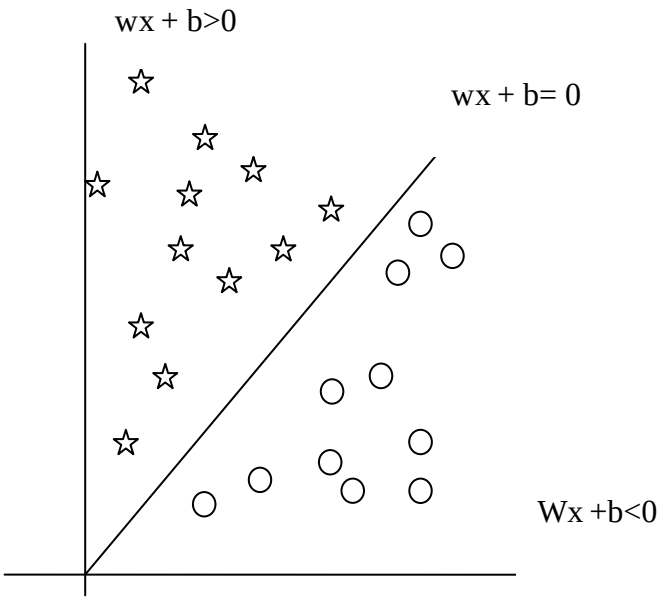


Figure 1.7 Linear SVM

II. Non-Linear SVM:- In the non-linear SVM plane is not linear. In this, the kernel function is used for the transformation of input space into high dimensional space as shown in Figure 1.9.

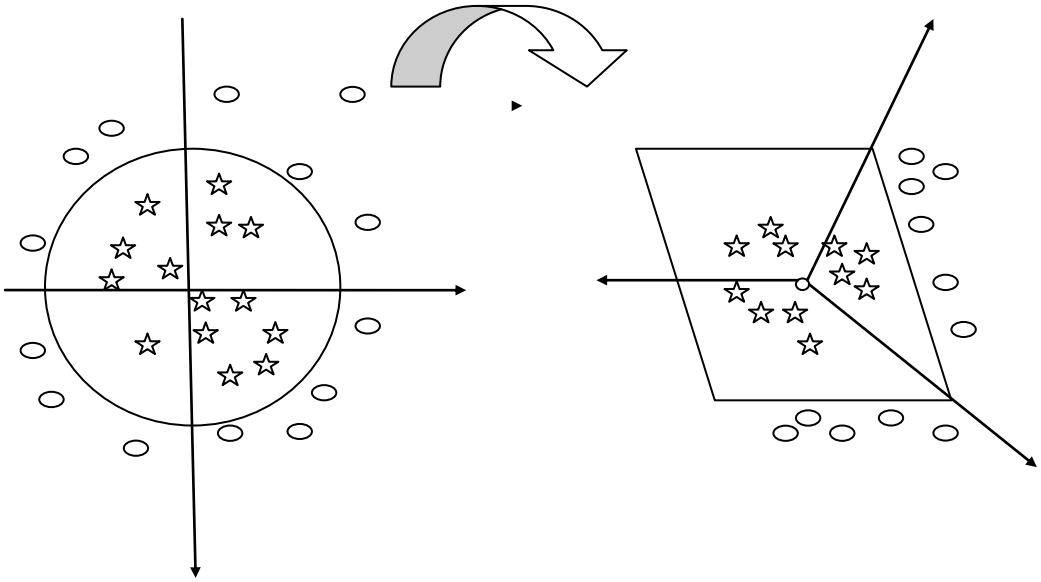


Figure 1.8 Non-Linear SVM

1.3 Clustering Technique

The process of grouping a set of physical or abstract objects into classes of related objects is called clustering. A cluster is a collection of data objects that are related to one another within the similar cluster and are unrelated to the objects in other clusters [9]. Clustering is used to discover the pattern information [15]. Clustering is an unsupervised machine learning technique whereas classification is a machine learning technique. In classification, we classify the data based on class label given to the data. In clustering class label is not known. Therefore clustering is known as unsupervised technique. There are various clustering techniques. The one of the most commonly used technique is K-Mediods. In our proposed work K-Mediods clustering technique is used.

1.3.1 K-Mediods

K-Mediod is a clustering algorithm .K-mediod algorithm similarly work as K-mean algorithm with the slight difference. K mean algorithm calculate the mean of data items as a centroid where as K-Mediod randomly select the data points called mediods. Median refers to data items having less average dissimilarity with other data items in the cluster. K-Mediod algorithm is better than K mean algorithm because it is vigorous to outliers. The algorithm of K-Mediod is given below. K-mediod algorithm minimised R squared error. The complexity of the K-Mediod is O(k(n-k)2).

1. Randomly select k data items as Mediod from n data items from given dataset.
2. Calculate the difference between K median and n data items and assign data items to the median which is closed to it.
3. Calculate the total swapping cost TCij for non-selected data items if and selected data items j.
4. If TCj < TCi, i is replaced by j.
5. Repeat step two and three until there is no data item left [16], [17].

1.4 Principal Component Analyses

Principal component analysis is the feature selection technique. It is used to reduce the dimensions of data and also reduce the computational complexity. Principal component analysis is ranking based technique. It calculates the eigenvectors and eigenvalues. Based on eigenvalues, principal component analysis order the variables of features from ascending to descending. The features for variables of data set having the

lowest then the value can be eliminated with minimum Data loss. Principal component analysis reduces the data dimensions having a large number of variables and also increases the accuracy. With the least number of variables, it is easy to calculate the results with lesser time and high accuracy. It means principal component analysis also helps to increase the computational speed of the algorithm. Principal component analysis also finds the related variable and provides the ranking to a related set of variables. The algorithm of principal component analysis works as follows.

1. Find the table of the input matrix.
2. Subtract the mean.
3. Compute the covariance matrix.
4. Compute the eigenvectors and eigenvalues of the covariance matrix.
5. Select components and forming a feature vector.
6. Drive the new data set [18].

CHAPTER 2

LITERATURE REVIEW

A.Saberi et al. (2007)

This paper present ensemble method to detect the phishing scams. Data mining classification algorithms such Naive Bayes, K-nearest neighbor and Poisson probabilistic theory and Naive Bayes are used to classifying spam and non-spam. The result of these classifiers is combined to get higher accuracy. Naive Bayes, k-nearest neighbor and Poisson algorithm separately provide accuracy of 88%, 87.5%, and 90.6% respectively. After combining these three techniques, it provides increased accuracy with 94.4%. The accuracy to detect the scams can be improved by using other techniques such as Neural Networks and SVM [19].

Durgesh K. Srivastav et al. (2009)

This paper presents the overview of SVM and the selection of a kernel function among the functions. SVM is a supervised machine learning techniques which can be applied to various kinds of data sets. Dimensions of data and Limited samples do not create any limitation in an SVM. There are two types of SVM, linear SVM, and nonlinear SVM. For nonlinear data set, we can use nonlinear SVM with its kernel functions. The commonly used kernels are the linear kernel, polynomial kernel, RBF kernel and sigmoid kernel. In RBF kernel there are less numerical difficulties. It nonlinearly maps data sample into higher dimensional space and it also has fewer hyperplanes than the polynomial kernel. SVM always gives better accuracy than other algorithms [20].

G.Magno et al. (2010)

This paper presents the problem to detecting spammers on twitter. In this dataset of twitter is collected and labeled the pre-classified spammer and non-spammer users. Then attributes are identified based on the social behavior of the user. In this paper supervised machine learning technique SVM is used to discover the spammers. Radial Basis Function (RBF) kernel of Nonlinear SVM is used classifies very complex data. Based on the ten attributes this technique differentiates the spammer and non-spammers. In this 70% of spammers and 96% of non-spammers are rightly recognized. The approach

of this paper is also able to detect spam as an alternative of spammers. The accuracy of detecting spam is 87.2% [21].

Tajunisha et al. (2011)

This paper presents a technique to improve the K-mean algorithm efficiency and accuracy by using a principal component analysis approach. Principal component analysis is used for reducing the data dimensions which helps to improve the accuracy and reduce the complexity of the algorithm. In this paper principal component analysis is applied at a first step to identify the centroid of k-mean and which helps to reduce the dimension of data. This heuristic approach reduces the distance between data points to the cluster. The principal component analysis calculates the eigenvalues and eigenvectors. The dimension of data can be reduced by eliminating the eigenvectors having the lowest eigenvalues. The results of this technique are accurate than the existing technique [18].

J. Ratkiewicz et al. (2011)

This paper present of Framework that detects dispersal of political misstatement. In this machine learning technique is used to detect the political misstatement spread by hackers on Twitter. This Framework combines the topological content based and crowdsourced features to evaluate the behavior of users. To classifiers, AdaBoost and SVM are used to produce the results. Both classifiers are used with the re-sampling and without re-sampling. The accuracy of AdaBoost and SVM without Re-sampling is 92.6% and 88.3% respectively. With re-sampling, the accuracy of SVM and AdaBoost is 95.6% and 96.4% [22].

M.Secchiero et al. (2012)

In this paper concept to detect the fake profile attacks is provided. This approach is based on increasing charge of the social network graph. For experiment analysis, the dataset is collected by using sensing application to get information of user profile on social networks. In this behavior of the user is identified to differentiate the real user and attacker. Attacker behaves differently and not involves those people which are close to the real user to avoid the detection. The increasing rate of people is also different than the real user. This leads to different behavior of an attacker. In this approach, the time evolution is called to raised an alert flag when test profile diverge from normal behavior [5].

S.Kiruthiga et al. (2014)

This paper presents the techniques to detect the cloning attacks in social networks. In this, Dataset of facebook is collected from the university of California for experiment analysis. In this paper, clone profiles are detecting by identifying the similarities between two profiles. Naive Bayes classifier is used to classify the information of users on facebook having maximum similarity. K-Mean clustering is used to group the users having same attributes such as same college etc. The similarity between real and clone profiles is detected by cosine similarity. To improve the similarity results clone spotter algorithm is used [23].

M.Alsaleh et al. (2014)

This paper focused on detecting the Sybil in twitter. A Twitter dataset of the user, Sybil and hybrid are collected and labeled. a variety of features is analyzed to detect the Sybil. Four classification techniques Decision Tree, Random Forest, SVM and Multilayer Neural Network is used to classify the data. But Multilayer Neural Network provides the more detection rate and lower error rate as compare to other classifiers. Detection rate and error rate in Multilayer Neural Network is 88.57% and 11.43% respectively. In this browser plug-in is also developed which is notify the user about Sybil accounts before accessing them [24].

Y.Shen et al. (2014)

This paper present binary classifier to detect the forged followers by derives the characteristics of fake followers in sina weibo. These features are divided into three parts- the post related features, user relationship features and evolutionary features. SVM is used as the main classifier. 10 fold cross-validations are done on the dataset. This approach provides the higher accuracy and lowers false positive than other approaches. The accuracy of this approach is 98.7% and false positive is 0.4% which is very near to 0 [25].

F. Michael et al. (2014)

This paper presents the solutions to protect from the threats which faced by users on social networks. First of all, this paper discusses various categories of threats which targets all users on social networks including children and youngsters. It divides the threats into four parts. The first kind of threats not only targets the social networking users but also those are not used social networking sites but simply on the internet. Second type threats only target the social networking users. In third type attacker

combines two or more threats and fourth category threat targets only children. According to the type of threats, this paper presents the three solutions- operator solution, commercial solution and academic solution to provide confidentiality and safety [26].

Sonali. B. Maind et al. (2014)

This paper presents the overview of neural networks working and training. It also provides the advantages and application of the neural networks. Neural Network works like a human brain. Neural Network has multiple neurons interconnected with each other. There are three layers of neurons input layer, an output layer, and hidden layer. The learning process of the neural network is similar to a human brain i.e it learns by examples. The neural network has many applications. The basic applications of the neural network are pattern recognition and classification of data. The neural network has a good performance and provides very good accuracy. The neural network is a best for real-time applications because it has a very good learning ability that is learning how to do a job based on given training dataset [7].

Gandhi Gopy et al. (2014)

This paper presents modified algorithm for k-Mediods. This paper also compares the results of modified K-mediod with the existing k-mediod and k-mean. The improved K-mediod algorithm works faster than the existing k-mean and k-mediod algorithms. k-mean works based on the mean value of data objects. It randomly selects objects which represent the mean or center of clusters and assign the other objects to the cluster by comparing their similarity. k-mediod similarly works as k-mean, it randomly select the objects to represent mediod and all objects are assigned to cluster with the nearest to the mediod. After that, it process all the data object and identified new mediod to represent a cluster in better way.k-mediod is very sensitive to outliers and this limitation is removed by modified k-mediod [27].

Kumar Parveen (2014)

This paper presents the study of neural network its characteristics and applications. The neural network is a complex structure which consists of multiple neurons which are interconnected. The neural network is a machine learning technique which is works as a human brain. The neural network is basically used for pattern recognition and classification. The neural network is used to solve the complex problem

and real-time operations. The neural network has three layers input layer output layer and hidden layer. The working of the neural network is based on these three layers. The neural network has learning ability about how to do a task. The neural network gives better accuracy based on training data set. The limitations of the neural network are many it does not describe how they solve the problem. It is also not used to solve the daily life problems [28].

M.Egele et al. (2015)

This paper presents a system named COMPA to detect the compromised accounts on social networks. This system is based on the behavior of users on social networks. The behavior of normal users are stable and COMPA detects compromised accounts having behavior more inconsistent. In COMPA behavioral profile is generated based on the previous message sent by the account. When a new message is created, the comparison is done with the behavioral profile. If the message is variant with a behavioral profile, COMPA flags it as a compromise. This technique is applied on both twitter and facebook and provides good results. The false positive for twitter and facebook is 4% and 3.6% respectively [29].

D.Freeman et al. (2015)

This paper focused on detecting the clusters of fake accounts rather than an individual. This approach created a cluster, based on the features provided at the registration time such as registration IP address and registration date. Random forest, SVM, and Logistic regression is used to train the model and SVM are used to classify the cluster of accounts as fake or not. This approach provides the fast detection of fake accounts. This approach is applied on the LinkedIn dataset and provides the 95% precision [30].

G. Supraja et al. (2015)

This paper presents the pattern detection approach to detect the fake accounts. In this paper, the crawler is used to collect the twitter dataset. The pattern matching algorithm is used on the screen name and updates time of tweets to detect a group of fake accounts. The time to create the profile is analyzed for detection. The time taken by the fake user is different than the real user. The advantage of this approach is that it is a fast approach to detect the fake accounts. The disadvantage of this approach is that it detects only fewer numbers of fake accounts [31].

Yazan Boshmaf (2015)

This paper presents an approach to thwart fake accounts in social networking sites. In this paper real-time data set for Facebook and tuenti is collected. This approach detects the fake accounts based on feature set. The 18 features of Facebook is extracted from facebook data set and 14 features extracted from twenty dataset. The random forest classifier is used to classify the data. In a random forest 450 decision trees are constructed at training time for Facebook data set and 500 decision trees are constructed for tuneti data set. The decision tree randomly selects three features of Facebook out of eighteen and seven features are selected from tuenti out of fourteen. Naive Bayes and SVM algorithms are also used. This approach use 10 fold cross validation method. The random forest gives the AUC of 0.7 for facebook and AUC of 0.76 tuenti. Navie Bayes gives AUC of 0.63 and SVM gives AUC of 0.57 for facebook. Naive Bayes and SVM gives AUC of 0.64 and 0.59 respectively for tuneti [32].

Rahman Sazzadur et al. (2015)

This paper presents the technique called FRAppE to detect the malicious applications on Facebook. FRAppE alert the user before installing the malicious application It is a feature based detection technique. To develop this technique data set is collected from facebook users. The feature set is identified which helps to detect fake applications on Facebook. Machine learning algorithm is used to classify the data. 5 fold cross validation are used to train and test the data in FRAppE. FRAppe detects the malicious application with the accuracy of 99.5% and the true positive rate is 95.9% [33].

Krishna B Kansara et al. (2016)

This paper proposed a Sybil node discovery method based on the social graph. This approach overcomes the limitations of the previous graph-based approaches by adding user behavioral manners such as latent dealings and friendship refusal. The proposed design is divided into two parts, Sybil node identification (SNI) and Sybil node identification using behavioral analysis (SNI-B). SNI method used classical Sybil detection and SNI-B is an extension of SNI. Results of both methods are compared. SNI-B provides higher accuracy, precision, and recall than SNI. The accuracy of SNI is 77% and 92% for SNI-B. The precision of SNI is 75% and 80% for SNI-B. Similarly, recall for SNI is 60% and for SNI-B is 80% [34].

A.Azab et al. (2016)

This paper presents the classification techniques to detect the fake accounts on twitter. To detect the fake accounts this paper used feature based approach. The minimum weighted feature set is used. In this approach, the behavior of the user is identified. The real user behaves differently than fake users. This behavior is used to identify the fake accounts. Different classification techniques such as random forest, decision tree, naive Bayes, neural network, and SVM are used. The accuracy of all techniques is provided. The gain measure is used to assign the weights to the feature set. To train and test the algorithms five-fold cross-validations are applied SVM gives best accuracy results to detect the fake accounts [6].

Ali M. Meligy (2017)

This paper presents a technique to detect fake accounts on social networking site called fake profile recognizer. This technique is based on two methods i.e regular expression and deterministic finite automata. A regular expression is used to authenticate the profiles and deterministic automata recognize the identities in trusted manner. This technique is applied on Facebook Google Plus and Twitter data set. The accuracy of Facebook, Twitter and Google + data set is 89.73%, 76.94%, 81.9% respectively. The Precision for Facebook is 88.9% for Google+is 77.41 Percent and for twitter is 81.81%. The false positive rate for Facebook is 11.66, for Google+ is 26.10% and for Twitter is 20.86%. The false negative rate for Facebook is 11.04%, google + 22.60% and for Twitter 18.20% [35].

Comparative Analysis

Comparison between various techniques and processing methods used in the different studies are given in Table 2.1. It shows the different classifiers and processing methods used to detect the fake accounts on social networking sites. The dataset link is also given if any.

Table 2.1 Comparative analysis

Author	Year	Classifier	Processing method	Dataset link
A.Saberi et al.	2007	Naive Bayes, K-nearest neighbor	Structural features	-
G.Magno et al.	2010	SVM	Social behaviour	http://twitter.mpi-sws.org
J.Ratkiewicz et al.	2011	Adaboost, SVM	Topological, content based	-
M.Secchiero et al.	2012	-	Social network graph	-
S.Kiruthiga et al.	2014	Naive Bayes	Cosine similarity	http://odysseas.calit2.uci.edu/research/data/hybrid/_releasenetworkmapping
M.Alsaleh et al.	2014	Decision tree, random forest, Navie bayes, multilayer neural networks	Feature set	-
Y.Shen et al.	2014	SVM	Feature set	-
F. Michael et al.	2014	-	-	-

M.Egele et al.	2015	COMPA	Behavioural profiles	-
D. Freeman et al.	2015	SVM, Random forest	Registration IP address	-
G. Supraja et al.	2015	-	Pattern Matching	-
Yazan Boshmaf	2015	Random forest, SVM, Naive Bayes	Feature set	-
Sazzadur Rahman et al.	2015	FRAppE	Feature set	-
Krishna B Kansara et al.	2015	-	Social graph	-
A.Azab et al.	2016	SVM, Neural Networks, Random forest, Decision Tree, Naive Bayes	Feature set	-
Ali M. Meligy et al.	2017	-	regular expression and deterministic finite automata	http://fastfollwerz.com

Parameters Used

There are different parameters are used in various studies to detect fake accounts in social networks. Those parameters and their values are given in Table2.2.

Table 2.2 Parameters used

Author (Year)	Parameters	Values
A.Saberi et al. (2007)	Accuracy	94.40%
G.Magno et al. (2010)	Accuracy	87.20%
J.Ratkiewicz et al. (2011)	Adaboost, SVM	95.6% , 96.4%
M.Secchiero et al. (2012)	-	-
S.Kiruthiga et al. (2014)	-	-
M.Alsaleh et al. (2014)	Detection rate, Error rete	88.57%, 11.44%
Y.Shen et al. (2014)	Accuracy, False positive	98.75, 0.4%
F. Michael et al. (2014)	-	-
M.Egele et al. (2015)	False positive	4%
D. Freeman et al. (2015)	Precision	95%
G. Supraja et al. (2015)	-	-
Yazan Boshmaf (2015)	AUC	0.7
Sazzadur Rahman et al. (2015)	Accuracy, True positive	99.5%, 95.9%
Krishna B Kansara et al. (2015)	Accuracy, precision, Recall	92%, 80%, 80%
A.Azab et al. (2016)	Precision, Recall, F-measure	99.5%, 74.76%, 85.40%
Ali M. Meligy et al. (2017)	Accuracy, Precision	89.73%, 88.9%

CHAPTER 3
PRESENT WORK

In on-line social networks to detect the fake accounts is a major challenge. The people using on-line social networks suffer from the various problems, which affects their personal as well as business life. The number of fake accounts on a social network is increased. Online social network suffers from fake accounts which are created. Fake accounts present fake news, web rating, and spam. Our proposed plan detects the fake accounts in the facebook. There are various techniques are available to detect the fake accounts on the on-line social networks. Each has their own advantages and purposes. But still, existing methods do not have a very high value of f-measure and recall value. This proposed work combines the weighted feature set with machine learning techniques to obtain the best results Using the proposed technique for detecting the fake accounts on facebook will improve the accuracy and exactness.

This proposed work uses the techniques like neural networks and support vector machine for classification of real and fake accounts. The feature set that influences the detection of fake accounts detection of the fake on Facebook will be used. This proposed work is expected to generate the higher value of f-measure and recall required for detection of fake account in facebook. The machine learning techniques are neural network and Support vector machine provides the accurate results. Neural network and Support vector machine gives the better results in data classification. Machine learning techniques have been widely used in promoter prediction techniques because of its capability to be taught and resolve many real time problems. They can adjust their inner configuration without human intervention to produce estimated outcome for the specified problem and to find a connection between input and output. Therefore neural network and support vector machine results in higher accuracy for detecting the fake accounts on facebook.

3.1 Problem Formulation

There are lots of problems on social networking sites; one of the problems is fake accounts in social networking sites which lead to various problems. It affects the users on social networking sites in many ways. Online social network suffers from various fake accounts. There are very fewer techniques to find the fake accounts on the facebook. Even existing methods do not have very high accuracy. This proposed work combines the

weighted feature set with machine learning techniques to obtain the best results. Neural network and Support vector machine have a capability to learn and solve the problems. These techniques are used to solve the real time problems and provide very accurate results.

3.2 Objectives

The objectives of the work proposed plan include:

(i) To define the scope of the field to detect the fake accounts on social networking sites.
(ii) To study the various machine learning techniques used for classification for data.
(iii) To study the features for detecting the fake accounts.
(iv) To identify the required and optimal techniques for desire results.
(v) To implement the proposed technique to detecting the fake accounts.
(vi) To find the results.

3.3 Research Methodology

In the proposed work we use a hybrid approach to detecting a fake account on Facebook. We combine Different techniques to produced higher accuracy.

The methodology consists of following steps.

1. Dataset Collection

The first step of detecting the fake account on Facebook is collect the data set of Facebook. For the proposed work the data set of the Facebook is collected by survey method. By using survey method we collect the data set of Facebook from Facebook users. For this purpose, we create a Google form. Google form consists of various types of questions which help us to accurately classify the data free accounts and fake accounts. Dataset is collected online by using google forms and manually filling the form by Facebook users. We collected the data of 500 accounts of Facebook. We extract following feature set from the collected dataset of Facebook. This feature set consists of 16 features which help to accurately classify the data.

I. Number of facebook friends.
II. Number of photos Shared on Facebook.
III. Number of status/news shared per month.
IV. Number of groups joined.

V. Number of likes made per day.
VI. Number of days since updated the profile.
VII. Year in which joined the Facebook.
VIII. Number of pages liked.
IX. Number of posts liked by a Facebook friend.
X. Account user profile photo.
XI. Account as a cover photo.
XII. Frequently used hashtags in posts.
XIII. Account is logged in using iPhone.
XIV. Account is logged in using an Android phone.
XV. Number of Facebook friend those tagged the user.
XVI. Number of Facebook friends tagged by user.

2. Filtration of Dataset

In the second step with the filter the collected data set for filtration we apply a randomized filtration technique. Randomization randomly changes the position of accounts in the dataset. The filtration is also used to filter the wrong values filled in the dataset and the wrong value is replaced with the average value of its upper and lower column value. Filtration filters the dataset to accurately classify the dataset. If dataset does not contain any wrong value or null value then classification algorithm correctly classify the dataset.

3. Clustering of Dataset

After the filtration clustering technique is applied to the data set. K-media clustering technique is applied to dataset set witch assign the data set to clusters. There are two clusters for a data set. Cluster 1 and cluster 2, cluster 1 contains the data of fake accounts and cluster 2 contains the data of real accounts. Clustering technique detects the multiple fake accounts at a time which increases the accuracy and reduces the time complexity. A cluster of fake accounts is identified by using clustering technique. The k-mediod clustering algorithm is better than K-mean clustering algorithm because K-mediod is robust to outliers.

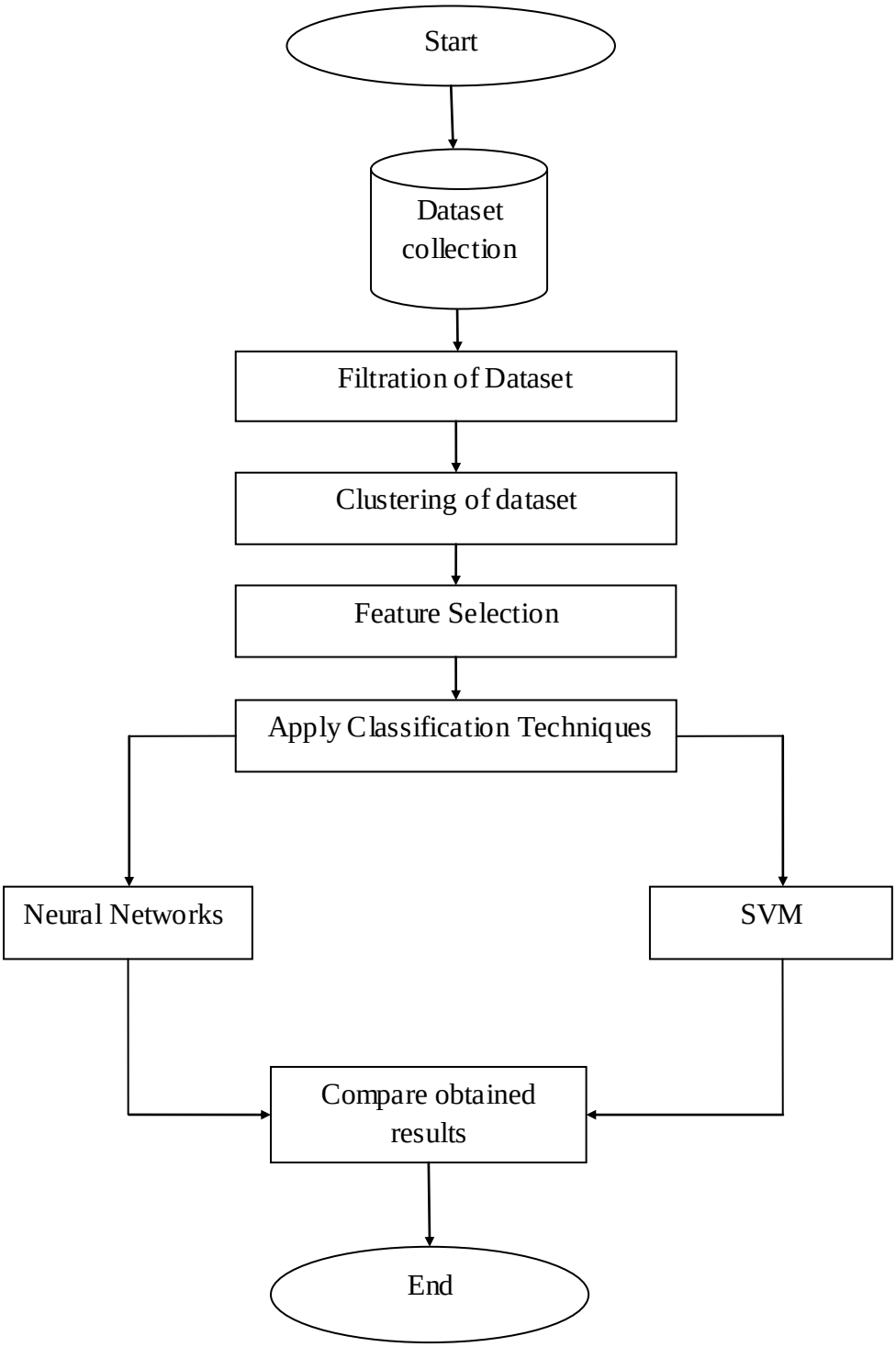


Figure 3.1 Methodology of Proposed Technique

4. Feature Selection

In this step feature selection technique is applied on the feature set. For feature selection principal component analysis is used for feature selection. Principal component analysis calculates the eigenvalues and eigenvectors and ordered them from higher to lower. Principal component analysis also combines the related features and assigns a value to them. Feature selection technique is used to get the higher accuracy with minimum features set. Because by using feature selection technique we can eliminate the features having the lowest weight zero weight. The ranking assigned to the attributes by principal component analysis is given below.

Table 3.1 Ranking of Attributes

S.NO	Ranking	Attributes
1	0.6423	Number of facebook friends, Number of friends you tagged, likes per day, Account has a cover photo.
2	0.5326	Number of groups you joined, pages liked, number of photos, Number of facebook user liked the post, frequently used hashtags in posts.
3	0.4555	In which year joined facebook, Number of new feed shared, pages liked, Number of groups joined, number of photos.
4	0.396	Number of days since updated the profile, In which year joined facebook, Frequently used hashtags in posts, Number of new feed shared, logged in an account using iphone.
5	0.3418	Number of facebook friends tagged the user, facebook friends tagged, logged in an account using iphone, number of new feed shared, number of post liked by facebook users, Number of friends you tagged.
6	0.2915	Logged in account using Android phone, logged in account using iphone, In which year joined facebook, Account has a cover photo, Account has a profile photo.
7	0.2461	Logged in your account using iphone, number of photos, logged in account using Android phone, Number of days since updated the profile, In which year joined facebook.
8	0.2053	Number of new feed shared, number of photos, logged in account using Android phone, In which year joined facebook, frequently used hashtags in posts.

9	0.168	Number of photos, number of new feed shared, logged in your account using iphone, number of facebook user liked your post, Number of groups joined.
10	0.1342	Number of facebook user liked your post, Number of groups joined, number of new feed shared, pages liked, logged in account using iphone.
11	0.1044	logged in account using iphone, frequently used hashtags in posts, account is connected with Instagram, Number of days since updated the profile, In which year joined facebook.
12	0.0779	Account has a profile photo, logged in account using Android phone, Account is connected with instagram, number of new feed shared, Account has a cover photo.
13	0.055	Account is connected with instagram, frequently used hashtags in posts, Account has a cover photo, logged in account using Android phone, pages liked.
14	0.0384	Number of likes per day, Number of facebook friends, Number of friends tagged, Account is connected with Instagram,

5. Classification of Dataset

Classification technique is applied after clustering and feature selection technique. Neural networks and SVM is used for classification of data. This two technique and most successful techniques and always give higher accuracy than other algorithms. Neural network and SVM are machine learning techniques which efficiently work on different kind of data sets. The execution time of SVM is less because it takes less time to train the SVM. The RBF kernel function is used in SVM. RBF kernel function performs better than polynomial kernel function and it is less complex. In our proposed work we use 10 cross-fold validation method for training and testing the data. The classification algorithms are evaluated and tested using 10-fold Cross Validation approach. In this, the training set is divided into 10 smaller sets and results are analyzed. 10-cross fold validation method divides the data set into 10 parts. 9 parts of dataset are used for training and one part is used for testing. Best training and testing process is repeated ten times with different training and testing data set.

6. Compare the Results

Two different results are produced by neural network and SVM in existing technique and proposed technique. The results of both techniques are compared with existing techniques. The parameters used and time taken by existing technique and proposed hybrid technique is compared. The results of Neural Network and SVM in proposed hybrid technique is also compared and the results with higher accuracy is considered.

CHAPTER 4
RESULTS AND DISCUSSIONS

In this proposed work the fake accounts will be detected by using machine learning techniques on Facebook. Two most successful techniques are used to classify the real and fake accounts. Feature set approach is used for the detection. The feature set is identified which influences the detection of fake accounts. In the proposed work we will find the most accurate results by using Neural Networks and SVM. These techniques accept the random data and provide most accurate results.

4.1 Experimental Results

This section presents the simulation results of the work done and the proposed approach. The simulation has been done in Java Net Beans and weka tool. Weka is a collection of machine learning calculations for information mining errands. The calculations can either be connected specifically to a dataset or called from your own Java code. Weka contains apparatuses for information pre-processing, classification, regression, clustering, association rules, and visualization. It is likewise appropriate for growing new machine learning plans. NetBeans is an open-source project dedicated to providing software development products (the NetBeans IDE and the NetBeans Platform) that address the needs of developers, users and the businesses who rely on NetBeans as a basis for their products; particularly, to enable them to develop these products quickly, efficiently and easily. The results are shown below in steps.

Results With Existing Techniques

1. Upload Dataset

   In the first window in graphical user interface, we upload the dataset of facebook. Dataset is saved in .csv format file. This interface helps us to choose the desired data set from any location and upload that data set. The data set which we upload comes under the files list The purpose of browsing is to select the dataset from any location within the computer so that it will be used for the further processing of the dataFrom the files we select the dataset and then whole details and data are loaded under the dataset. Figure 4.1 shows the screen to display dataset.

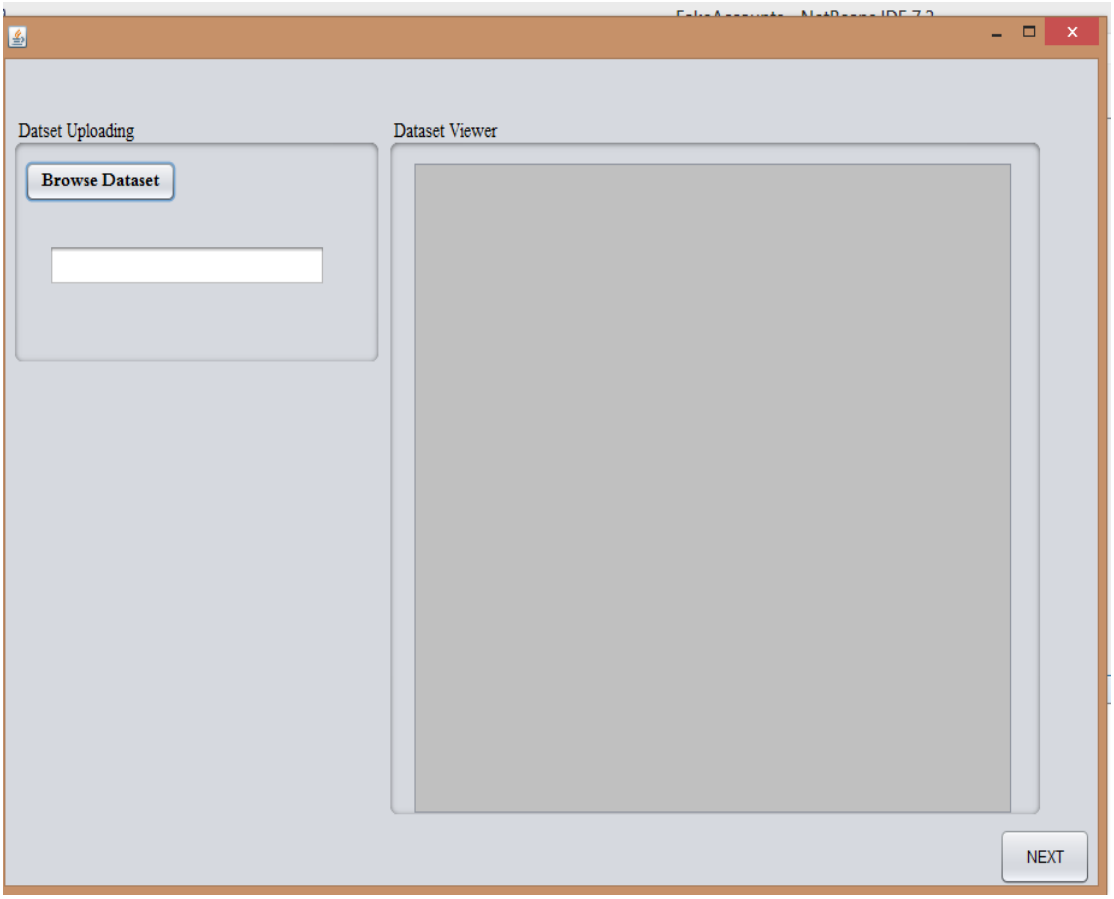


Figure 4.1 Browse Dataset

2. Choose dataset

In this step choose a dataset from a particular location. Dataset will be shown in the dataset viewer screen in figure 4.2. The dataset should be in CSV file. The CSV file of the dataset is selected from the computer drive where it is located. Using browse dataset button from user interface we select the dataset from a particular location as shown in figure 4.2. Figure 4.3 shows the selected dataset of facebook. It shows all the columns of CSV files with a class label. It displays the dataset of 500 accounts of facebook which are stored in CSV file. The Figure 4.4 visualization of dataset class label and feature set. This figure shows that all the features in the data set the visualization of class labels. It shows are 17 features at which are used to classify the real and fake account on Facebook. Based on this pictures at we classify the fake and real account on Facebook. Figure 4.5 shows the visualization of facebook dataset into weka tool.

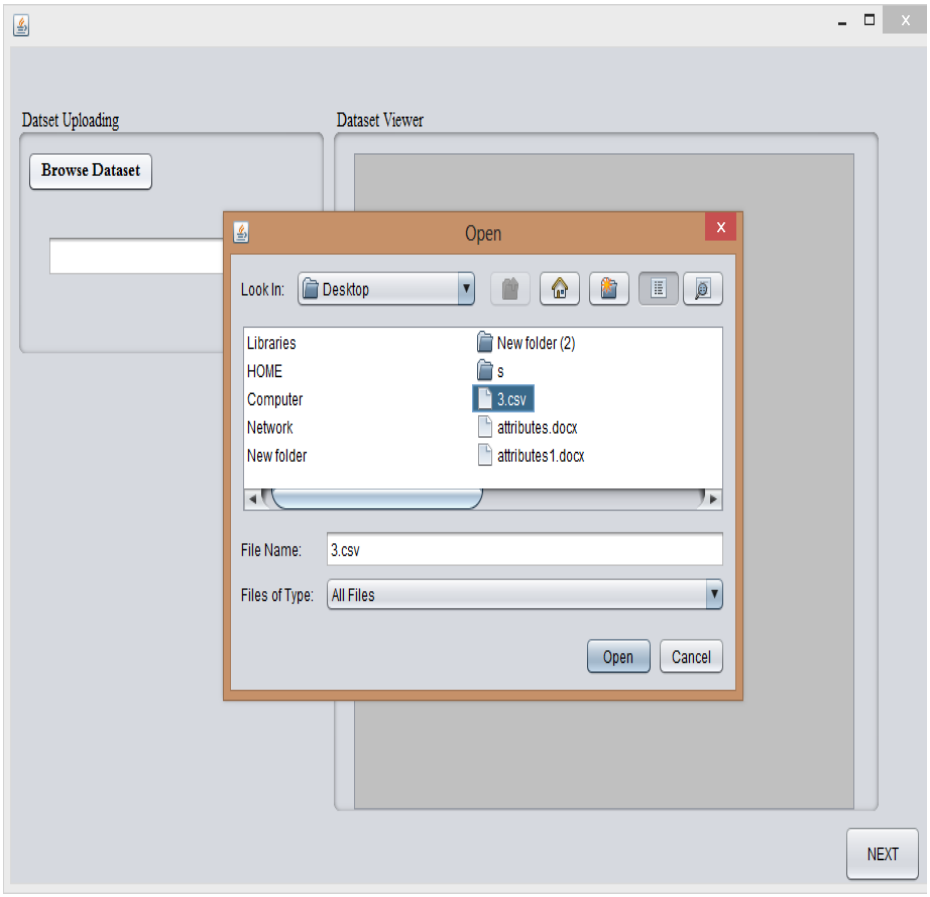


Figure 4.2 Choose Facebook Dataset

Figure 4.3 displays the selected dataset of facebook accounts. The dataset contains the data of 500 users of facebook.The data set contains these features: Number of facebook friends, Number of photos Shared on Facebook, Number of status/news shared per month, Number of groups joined, Number of likes made per day, Number of days since updated the profile, Year in which joined the Facebook, Number of pages liked, Number of posts liked by Facebook friend, Account user profile photo, Account as a cover photo, Frequently used hashtags in posts, Account is logged in using iPhone, Account is logged in using Android phone, Number of Facebook friend that tagged the user. A number of Facebook friends tagged by the user. Based on these features the classification algorithm classifies the real and fake accounts. The each account also has a class label which is used in the classification of a dataset. The class label is required for classification of data.

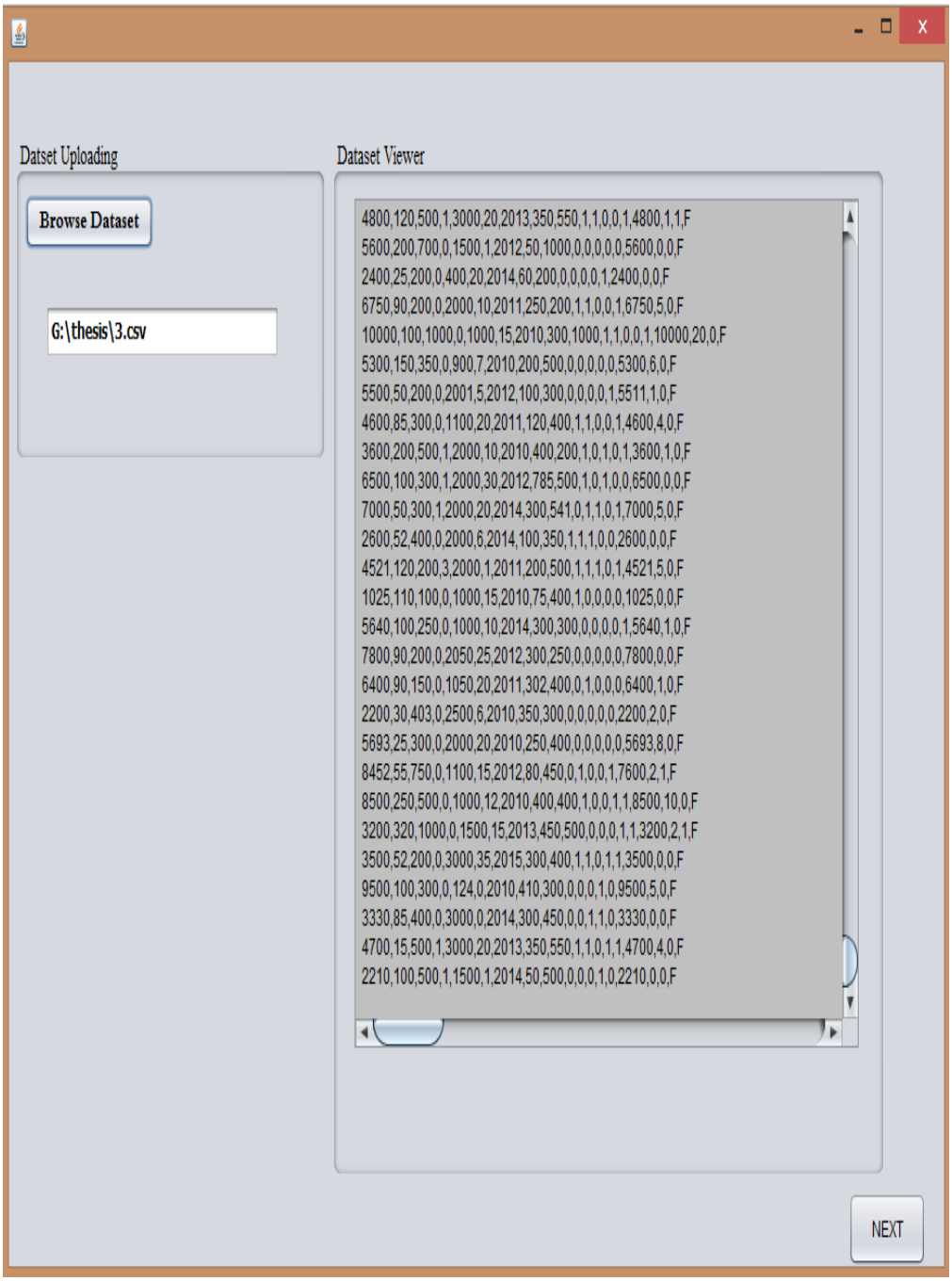


Figure 4.3 Facebook Dataset

Figure 4.4 shows the dataset selection in weka tool. All the features are displayed and also visualize the class label that is real accounts and fake accounts in the dataset. Visualization of class label and feature set in weka tool is given in figure 4.4. It shows 351 accounts are of reals users in Facebook and other accounts are of fake users in Facebook. It can be predicted from class label of dataset. The feature set contains 17 features in dataset and one class label for each account.

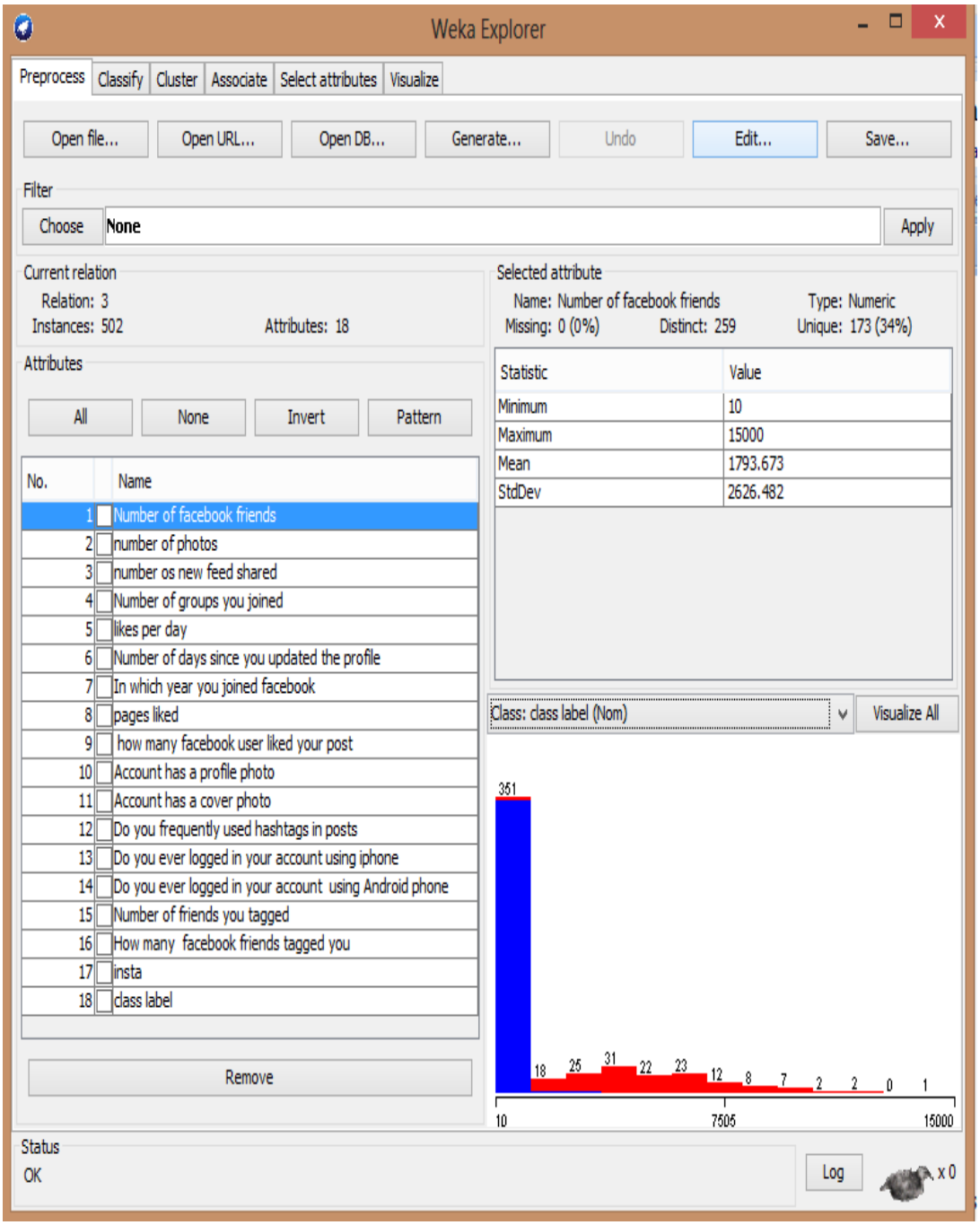


Figure 4.4 Feature set and class label visualization of dataset in weka tool

The visualization of the dataset in weka tool is shown in figure 4.5. It shows the visualization of each feature set in the dataset. This visualization is based on the class label of each feature set. It separately shows the visualization of each feature set in the dataset. The features are Number of facebook friends. Number of photos Shared on Facebook, Number of status/news shared per month, Number of groups joined, Number of likes made per day, Number of days since updated the profile, Year in which joined

the Facebook, Number of pages liked, Number of posts liked by a Facebook friend, Account user profile photo, Account as a cover photo, Frequently used hashtags in posts, Account is logged in using iPhone, Account is logged in using an Android phone, Number of Facebook friend those tagged the user, Number of Facebook friends tagged by user. The visualization of this feature set and class label is given in figure 4.5

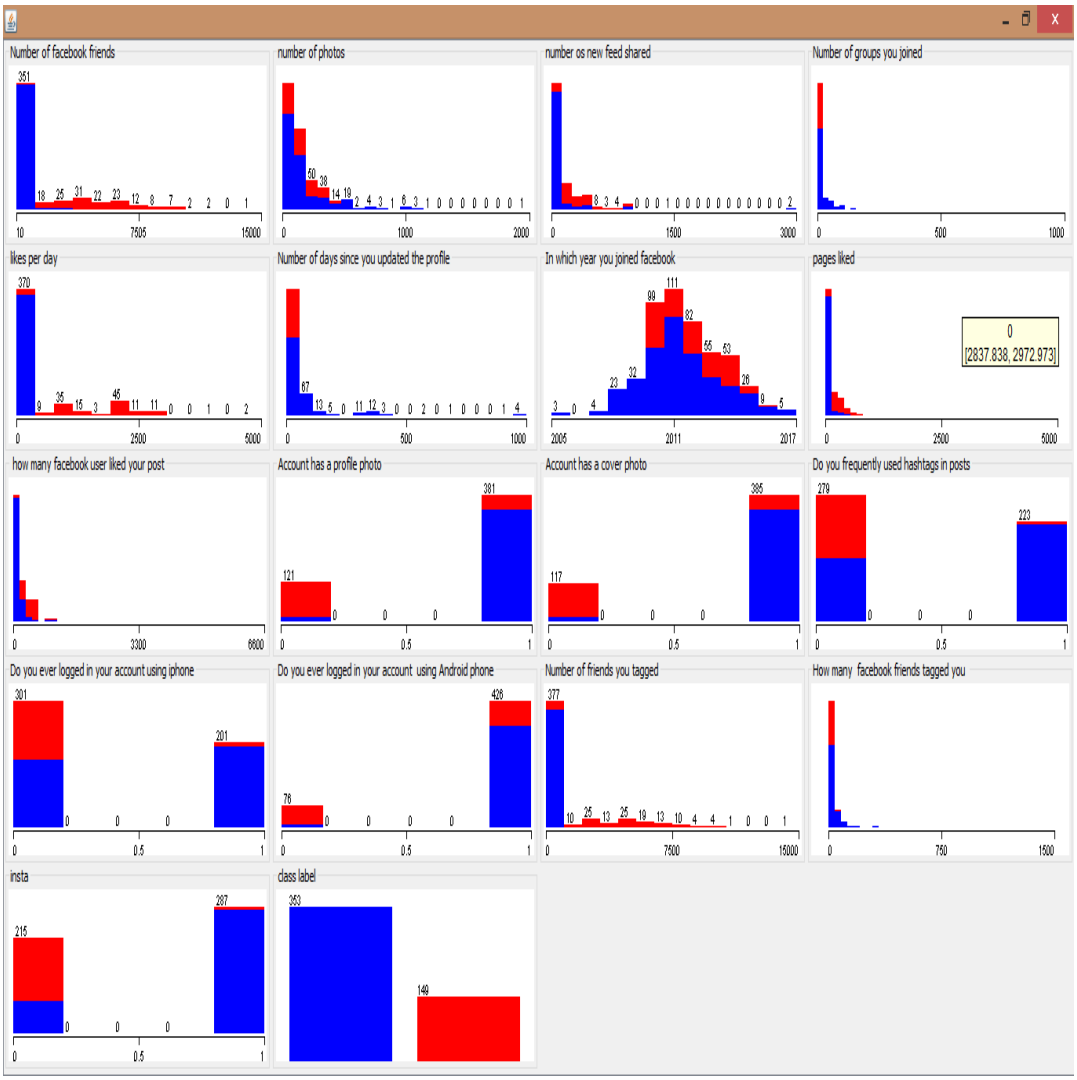


Figure 4.5 Visualization of Dataset in Weka Tool

3. Dataset Filtration

To filter the dataset randomization technique is used. Randomization randomly changes the position of accounts in the dataset. The filtration is also used to filter the wrong values filled in the dataset and the wrong value is replaced with the average

value of its upper and lower column value. Filtration filter the dataset to accurately classify the dataset. Figure 4.6 shows the filtration of the dataset. if dataset does not contain any wrong value or null value then classification algorithm correctly classify the dataset.

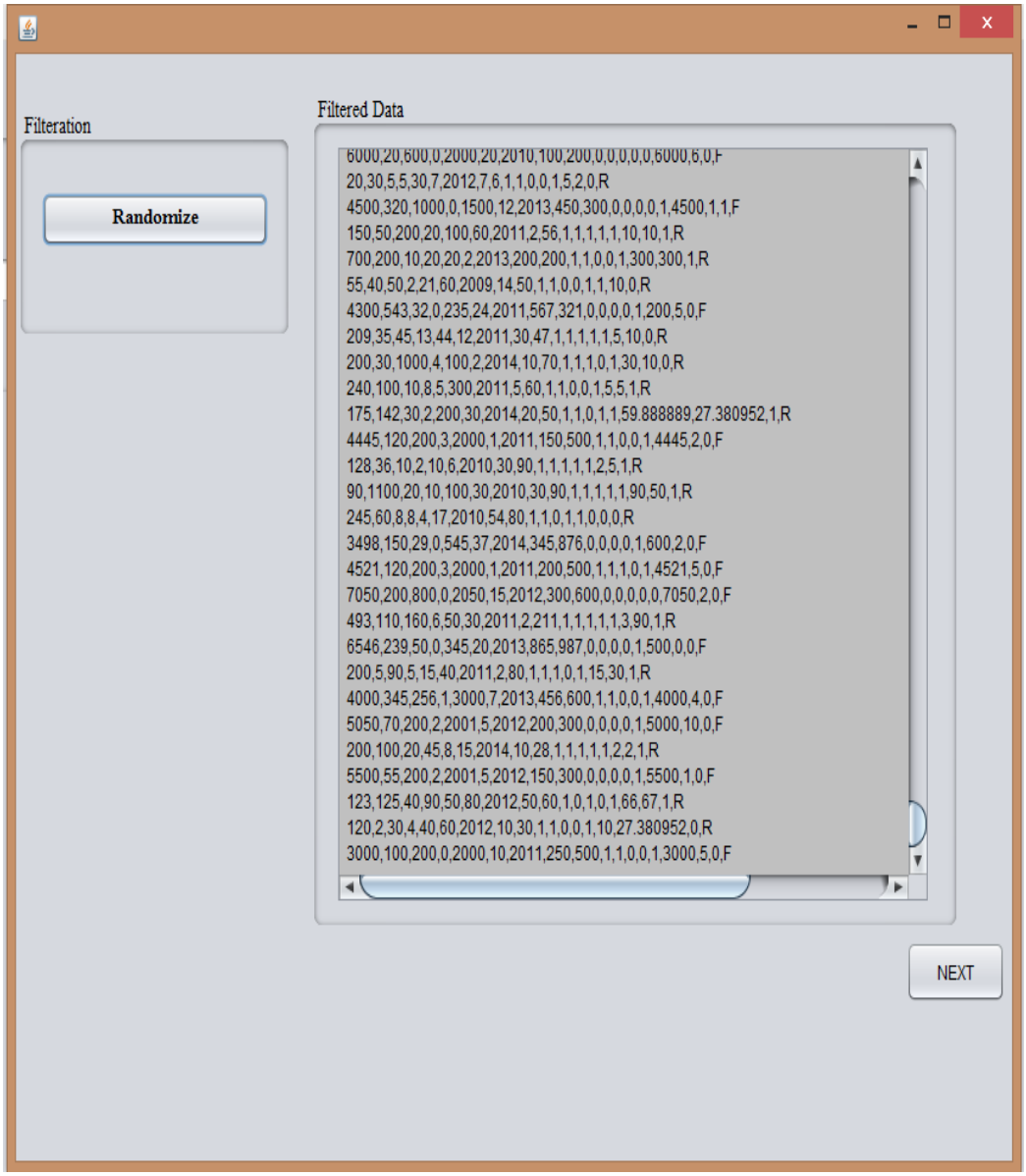


Figure 4.6 Filtration of Dataset

4. Classification with Neural Networks

Figure 4.7 shows the results of existing Neural Networks. The neural network in existing technique correctly classified 495 instances and incorrectly classifies 5 instances. It provides accuracy of 98.80%, the precision of 98.81%, recall is

98.80%and f-measure is also 98.80%. The TP rate and FP rate is 98.8% and 01.3% respectively.

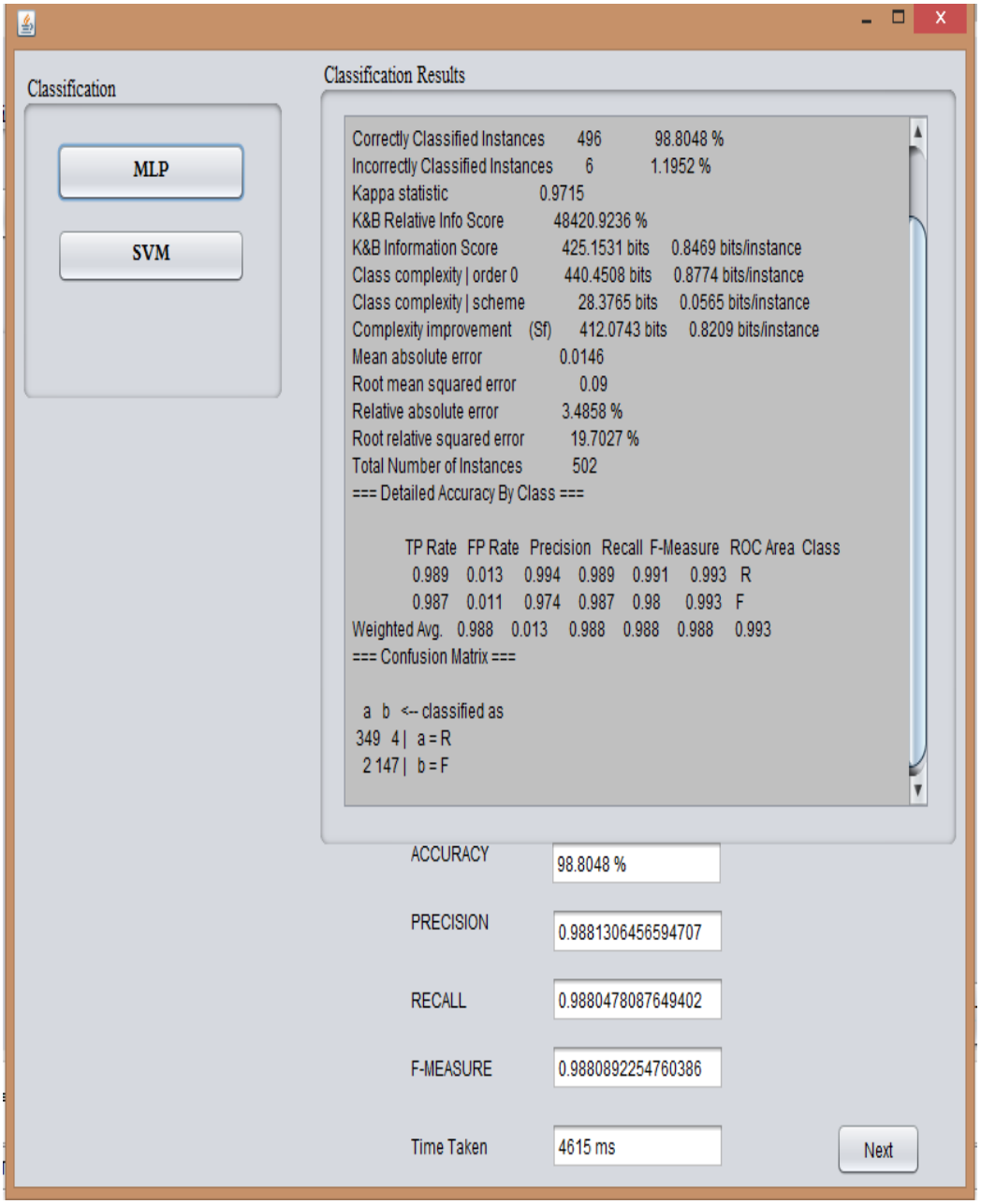


Figure 4.7 Results of Existing Neural Networks

5. Classification with SVM

Figure 4.8 shows the classification results of existing SVM. The result shows the accuracy of 71.11%. It correctly classifies the 357 instances and incorrectly classifies the 145 instances. It shows precision 79.52%, Recall 71.11%, F-measure 75.08%, TP Rate 71.1% and FP Rate 68.4%. The SVM provides very less accuracy in existing

technique because SVM in existing technique incorrectly classifies more number of instances. It is improved in proposed hybrid technique.

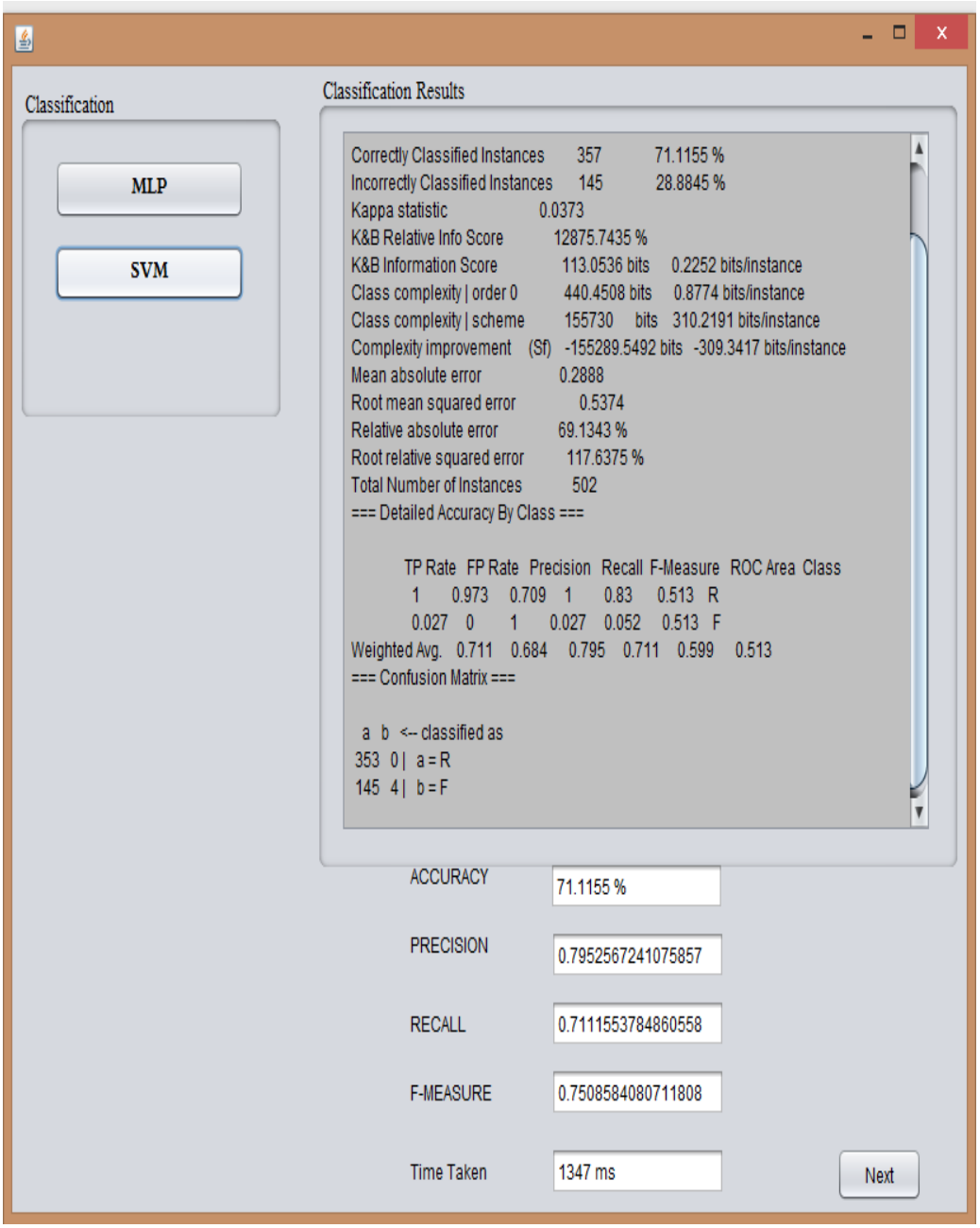


Figure 4.8 Results of Existing SVM

Results with proposed hybrid technique

6. Clustering of Dataset

Figure 4.9 shows the clusters of the dataset. The k-mediod clustering algorithm is used make the clusters of the dataset. This algorithm divides the dataset into two

clusters. Cluster 1 consists of fake accounts in dataset and cluster 2 consist of real accounts in the dataset. Clustering detects multiple fake accounts on facebook and improves the accuracy of proposed technique. Clustering technique detects the cluster of fake accounts at a time which also reduce the time complexity.

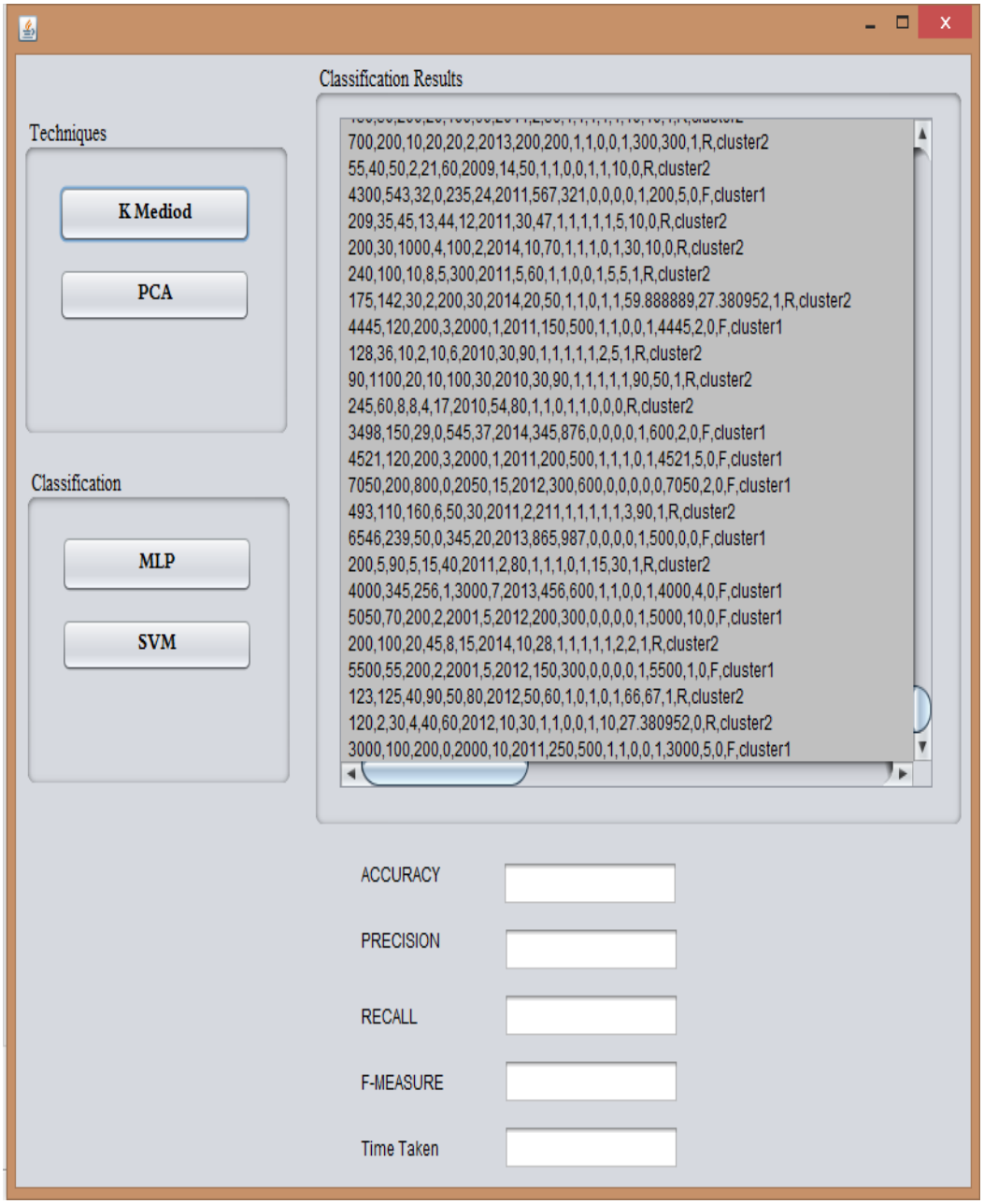


Figure 4.9 Results of clustering

7. Feature Selection

Principal component analysis feature selection technique is applied on the dataset. It is ranking based feature selection technique. In this dataset, it selects 14 feature set

out of 17 as shown in the figure 4.10. Principle component analysis provides a ranking to the feature set ordered from higher to lower.

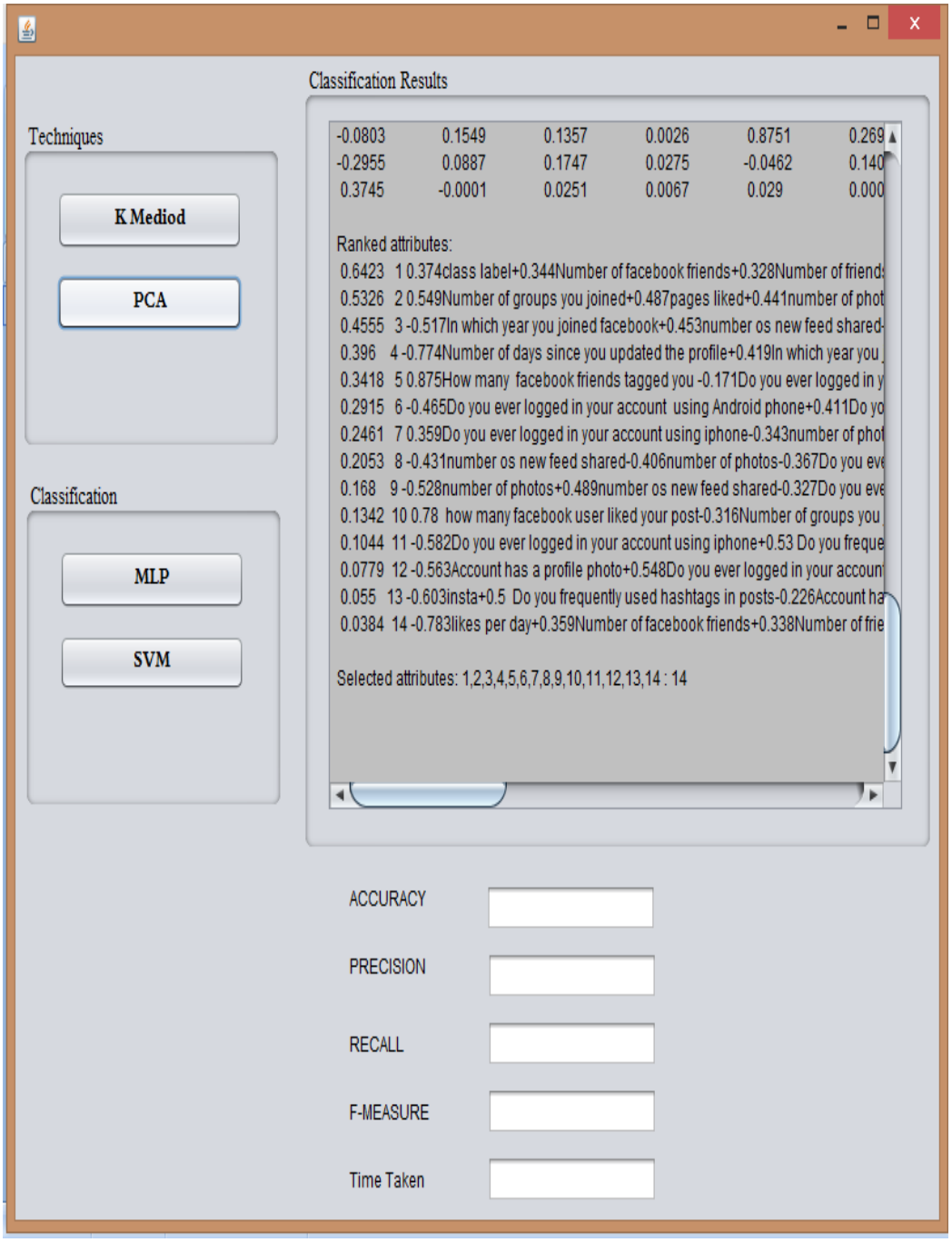


Figure 4.10 Results of Feature Selection

8. Classification with Neural Networks

In proposed hybrid technique Neural Networks shows the results as shown in figure 4.11. It shows the accuracy 99.40%, Precision 99.40%, Recall 99.40%, F-measure 99.40%, TP Rate is 99.4% and FP Rate 01.1%. It correctly classifies the 499

instances and incorrectly classifies only 3 instances. In proposed technique, neural networks execution time is 3548 ms which is less than existing technique.

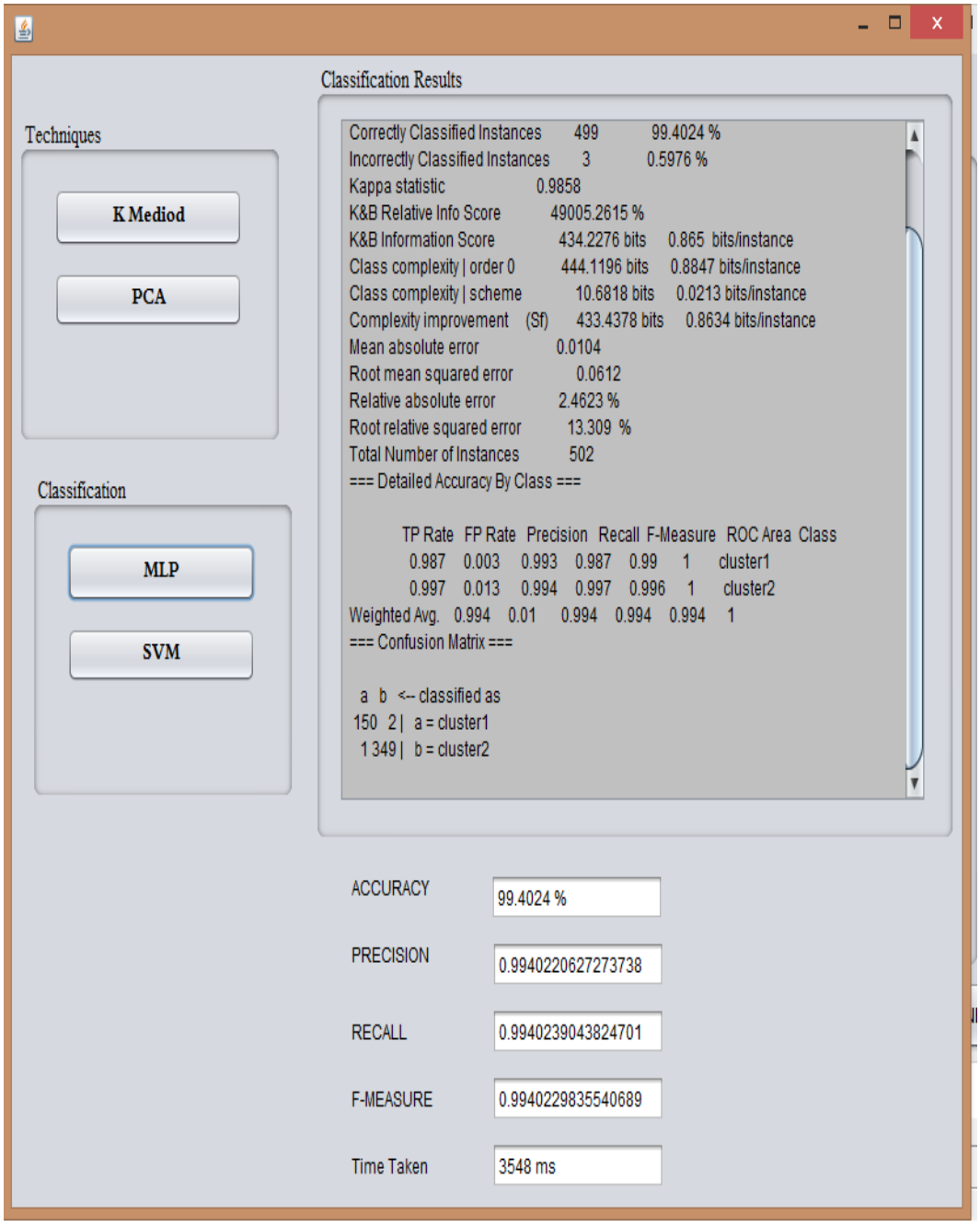


Figure 4.11 Results of proposed Neural Networks

9. Classification with SVM

In proposed hybrid technique the classification results of SVM is shown in figure 4.12. SVM correctly classifies 498 instances and incorrectly classifies 4 instances with 99.20% accuracy. The other parameters like Precision, Recall, F-measure, TP

Rate, FP Rate are 99.21%, 99.20%, 99.20%, 99.2%, 01.8% respectively. The time take by SVM in proposed technique is 471 ms. The execution time of SVM in proposed technique is very much less than the Existing technique. The proposed technique provides the higher accuracy with less execution time as compare to the existing technique. The accuracy of SVM in proposed technique is increased with Principle component analysis feature selection technique and execution time is also reduced.

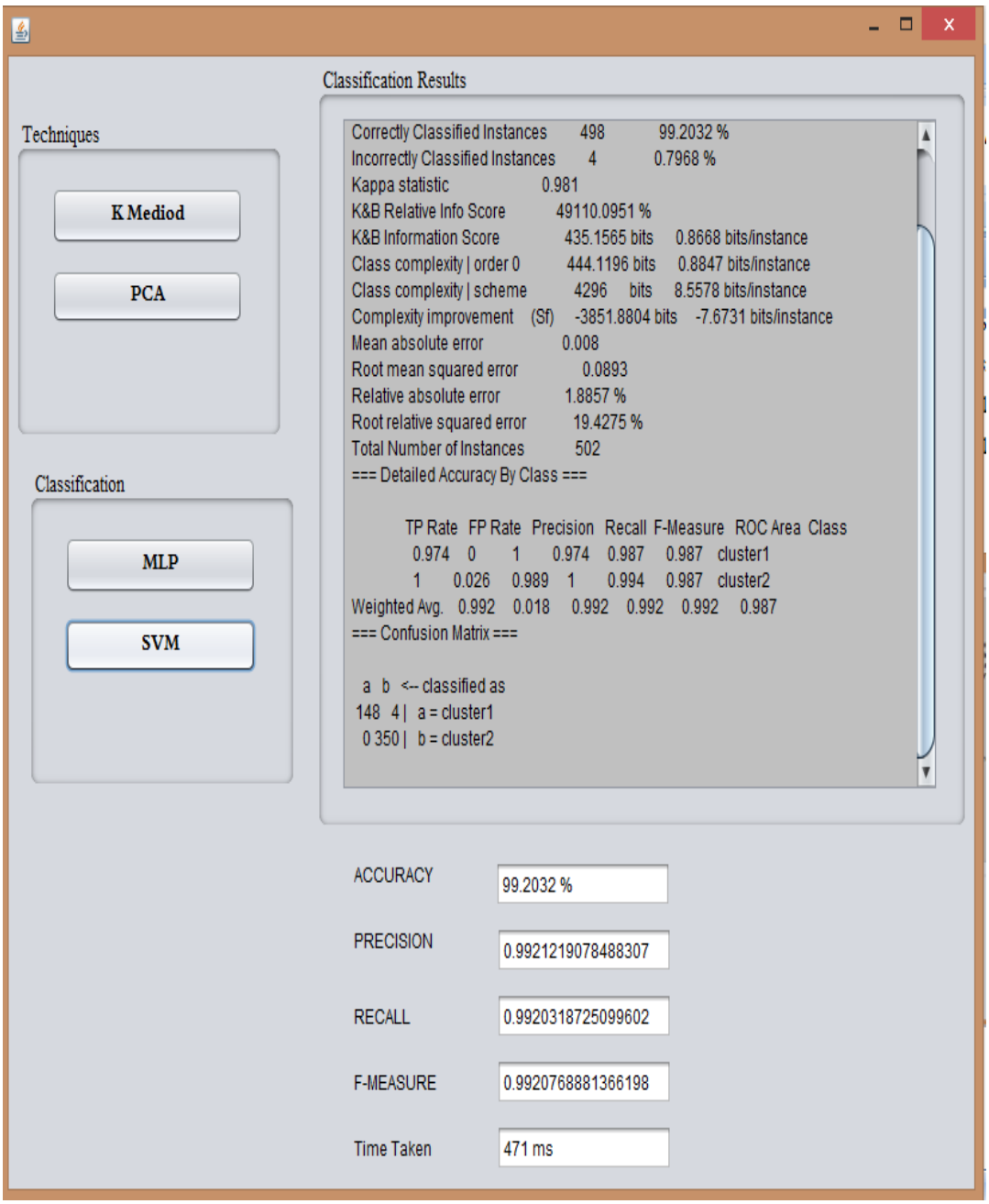


Figure 4.12 Results with Proposed SVM

4.2 Comparison with Existing Technique

The comparison between existing and proposed hybrid technique is shown with the various parameters. These parameters show that the proposed hybrid technique performs better than existing techniques. The comparison of Neural Networks and SVM in proposed technique is also compared. The parameters used to compare thre results of both techniques are Accuracy, TP rate, FP rate, Precision, Recall, F-measure and Execution time. There are following parameters are used to compare proposed hybrid technique with an existing technique.

1. **Sensitivity or TP Rate**

   It checks how many accounts have correctly classified a fake. It can be measured by a number of accounts that are correctly classified as fake to the total number of fake accounts. TP Rate should be high. Figure 4.13 shows TP Rate of existing Technique with the Proposed Technique. The TP Rate of existing Neural Network is 0.988 whereas the TP rate of neural networks in proposed technique is 0.994. Similarly thr TP rate of existing SVM is 0.711 and TP rate of SVM in proposed technique is 0.992. The proposed technique correctly identifies 498 accounts out of 502. It shows that the TP rate of proposed technique is higher than the existing technique.

2. **FP Rate (Fall-out)**

   It checks how many emails are incorrectly classified as spam. It can be measured as a number of emails that are incorrectly classified as spam to the total number of ham instances. It should be low. The FP rate of the existing and proposed technique is also shown in figure 4.13. The FP rate should be low. The FP rate of existing neural network is 0.13 whereas the FP rate of neural networks in proposed technique is 0.01. the FP rate of existing SVM is 0.684 and the FP rate of SVM in proposed technique is 0.018. The proposed technique incorrectly identifies only 4 accounts out of 502. The bar chart in 4.13 shows that the FP rate of proposed technique is lower than the FP rate of existing technique.

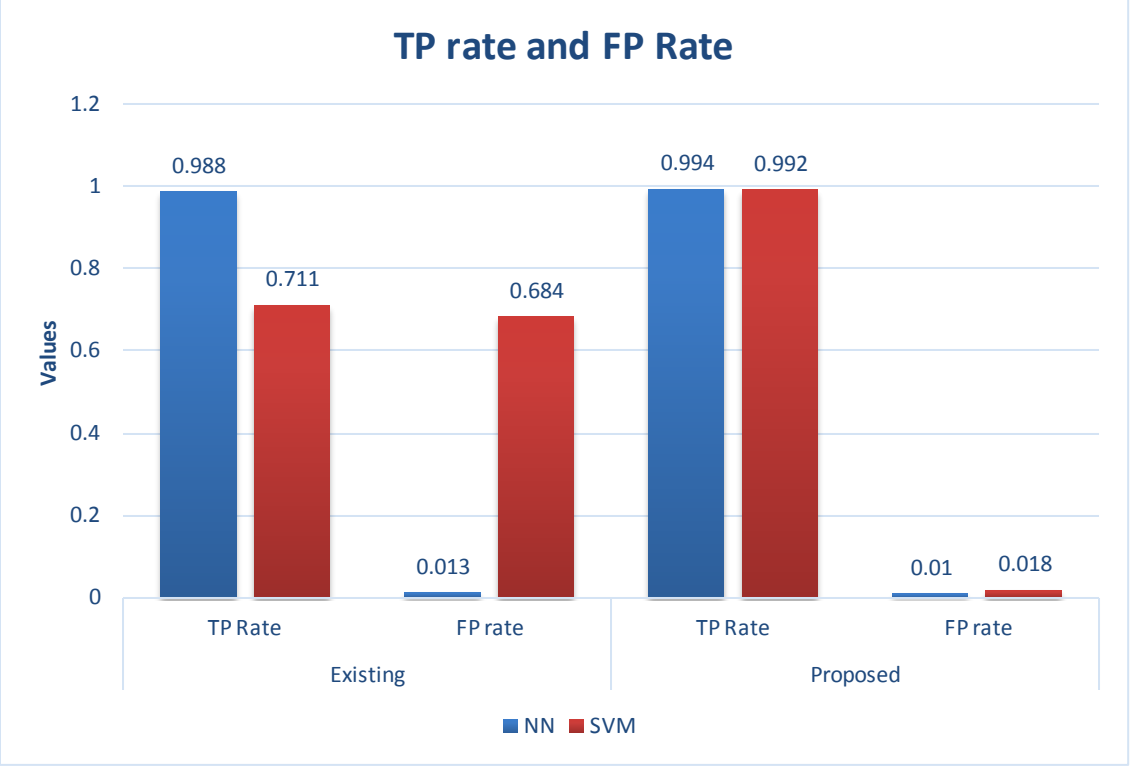


Figure 4.13 Comparison based on TP Rate and FP Rate

3. Accuracy

Accuracy is the percentage of correctly identified fake accounts. It can be measured as the number of correctly classified accounts to the total number of accounts. It should be highest for the best technique. The figure 4.14 shows the accuracy of the existing and proposed technique. The accuracy of proposed technique is higher than existing technique. The accuracy of existing neural networks is 98.80% and SVM is 71.11%. The accuracy of neural networks and SVM with hybrid proposed technique is 99.40 and 99.20 respectively.

$$\text{Accuracy} = \frac{\text{TP Rate} + \text{TN Rate}}{\text{TP Rate} + \text{TN Rate} + \text{FP Rate} + \text{FN Rate}}$$

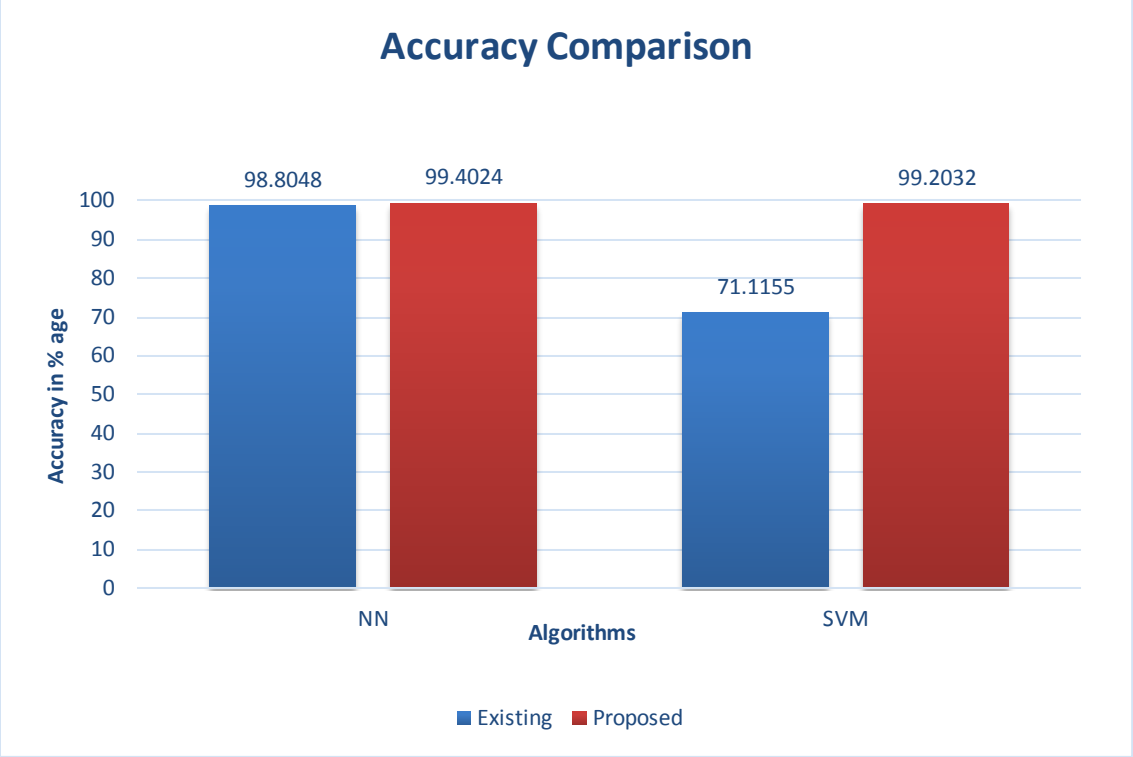


Figure 4.14 Comparison based on Accuracy

4. Precision

It checks how many accounts have correctly classified a fake among those all that are classified as fake. It can be measured by a number of accounts that are correctly classified as fake to the total number of accounts classified as fake. Figure 4.15 shows the comparison of an existing technique and proposed technique based on precision also. The precision should be high. The precision of existing neural networks and SVM is 0.9881 and 0.7952 respectively. The precision of neural networks and SVM in proposed technique is 0.994 and 0.992 respectively. The graph shows that the precision of proposed technique is higher than existing technique. The Precision is calculated as follows.

$$\text{Precision} = \frac{\text{TP Rate}}{\text{TP Rate} + \text{FP Rate}}$$

5. Recall

Percentage of correct documents that are selected in class from the entire document actually belonging to class. Recall should be high. The figure 4.15 shows the comparison between existing technique and purpose technique based on recall. The recall of existing neural network is 0.988 and existing SVM is 0.7111.

The record of proposed hybrid technique is 0.994 and 0.992 respectively. The figure shows that the recall proposes the hybrid technique is higher than existing technique. The recall is calculated by using true positive rate and false negative rate.

$$\text{Recall} = \frac{\text{TP Rate}}{\text{TP Rate} + \text{FN Rate}}$$

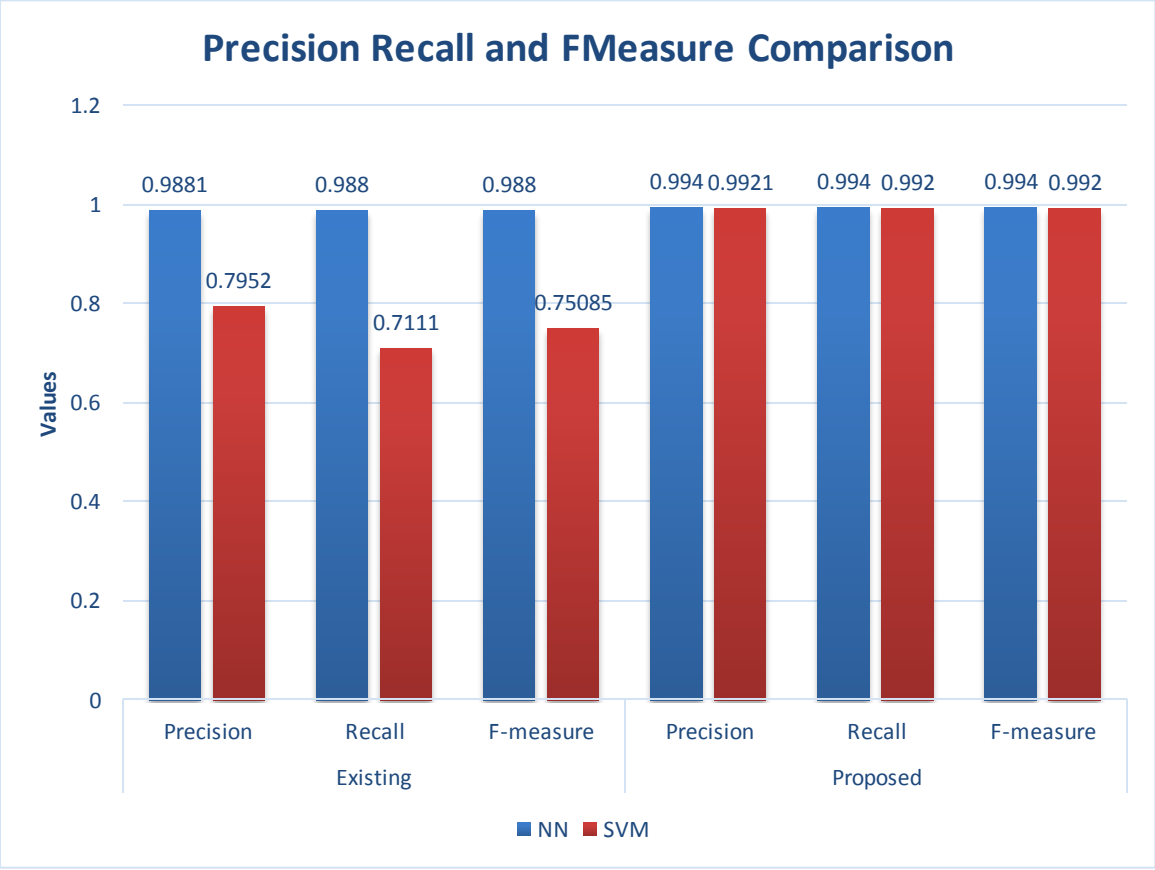


Figure 14.15 Comparison based on Precision, Recall and F-measure

1. F-Measure

It is defined as the weighted harmonic mean of Precision and Recall. It is also known as F-score. F-measure calculated from Precision and recall. F-measure also should be high. The figure 4.15 also shows the comparison between f-measure of an existing technique and proposed technique. F-Measure of neural network and SVM in existing technique is 0.9888 and 0.7508 respectively. The f-measure of neural

network and SVM in supposing the hybrid technique is 0.994 and 0.992 respectively, which is higher than the existing technique. The Precision of SVM in proposed technique is very much higher than the precision of existing technique. The graph shows that the f-measure of proposed hybrid technique is higher than the existing technique. The f-measure is calculated as follows.

$$F - Measure = 2 * \frac{Precision * Recall}{Precision + Recall}$$

2. Execution Time

It is defined as the time taken by the algorithms to classify the fake and real accounts. It should be less as possible. The figure 4.16 shows time comparison between existing technique and proposed technique. The time taken by both techniques is different in existing technique and proposed technique. Execution time should be less. The time taken by a neural network in the existing technique is 4615 ms and time taken by SVM in existing technique is 1347 ms. The time taken by neural network and SVM to classify the data is 3548 ms and 471 ms respectively. The graph shows that the time taken by existing technique to classify the data is much more than proposed technique. The hybrid technique takes very much less time to classify the data then existing technique. It means proposed technique also reduce the time complexity.

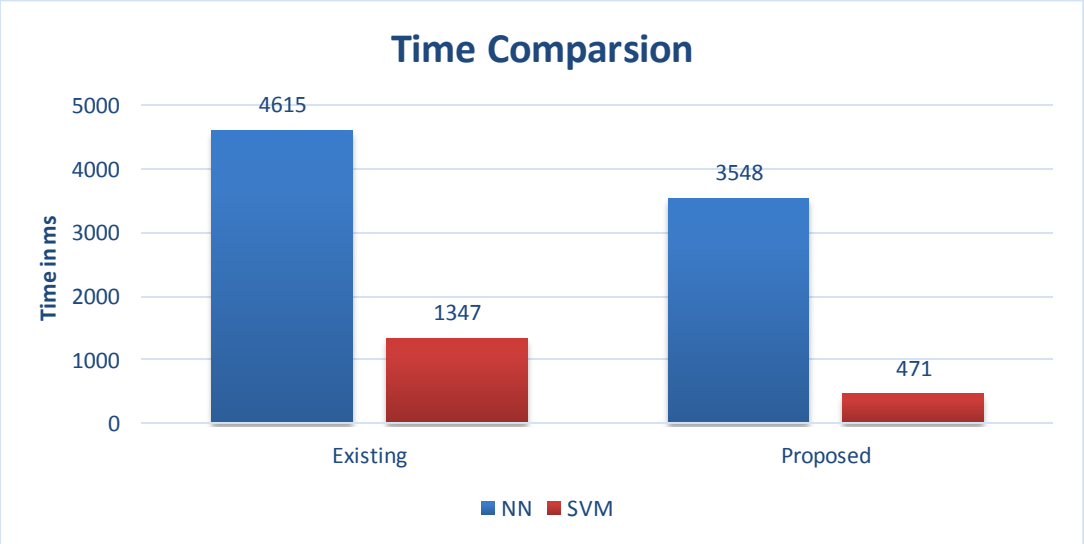


Figure 14.16 Comparison based on Execution Time

Table 4.1 shows the comparison between both techniques based on various parameters. It shows that the proposed technique is an improved technique the existing technique. The proposed technique provides higher accuracy with less execution time than existing technique.

Table 4.1 Comparison between existing and proposed technique

Parameters	Existing Technique		Proposed Technique	
	NN	SVM	NN	SVM
Accuracy	98.80	71.1155	99.40	99.20
Precision	0.9881	0.7952	0.9940	0.9921
Recall	0.9880	0.7111	0.9940	0.9920
F-measure	0.9880	0.7508	0.9940	0.9920
TP rate	0.9888	0.7111	0.9940	0.9920
FP rate	0.0130	0.6840	0.0100	0.0180
Execution time	4615 ms	1347 ms	3548 ms	471 ms

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

In the proposed was the hybrid technique is used to detect the fake accounts on facebook. In this proposed hybrid technique be used to most successful classifier neural network and SVM. K-mediod clustering is also used to improve the accuracy and reduce the time complexity of the algorithm. In proposed work collected real-time data set of Facebook from Facebook users. Then randomisation technique is used for data filtration. After the filtration of data set we apply K median clustering technique on the data set. The clustering technique assigns the data set to the clusters. There are two clusters of our data set one cluster for real account and second cluster for fake accounts. Clustering technique detects multiple fake accounts at a time. Clustering technique not only improves the accuracy to classify the data but also reduce that time complexity. Principal component analysis is used to provide the ranking on a feature set. It is a feature selection technique which provides the ranking to features higher to lower order. Using principal component analysis we can select those features which help to classify the data in the better way. Principal component analysis also used to increase the accuracy and decrease the time complexity by reducing the dimensions of data set. At the last step, we apply a classification technique neural network and SVM which classifies the data Neural network and testing machine learning techniques to classify the data. These techniques always give the better accuracy and other algorithms. In this purpose, we use RBF Kernel of SVM technique. The 10 cross fold validations are applied for training and testing the dataset. The 10-cross fold validation method divides the dataset into 10 equal parts and analyze the results.The accuracy of proposed work with neural network and SBM is 99.4% and 99.2% respectively. The Precision of proposed technique with the neural network is 99.4% and with SVM is 99.2%.

5.2 Future Scope

This proposed work presents a hybrid approach to detects the fake accounts on Facebook. In this proposed technique clustering, classification and feature selection algorithms are applied to get better results. The following points mention some idea that can be further implemented.

i. This technique can also be used for other social networking sites such as Twitter and LinkedIn with the minor changes.

ii. The accuracy of proposed technique can also be improved using different feature selection techniques.

REFERENCES

[1] G. Stringhini, "*Detecting Spammers on Social Networks*," *ACSAC*, pp. 1–9, 2010.

[2] I. Bara, C. J. Fung, and T. Dinh, "*Enhancing Twitter Spam Accounts Discovery Using Cross-Account Pattern Mining*," *IEEE*, 2015.

[3] Q. Cao, M. Sirivianos, X. Yang, and K. Munagala, "*Combating Friend Spam Using Social Rejections*," *IEEE*, 2015.

[4] S. Y. Wani, "*Prediction of Fake Profiles on Facebook using Supervised Machine Learning Techniques-A Theoretical Model* .," *IJCSIT*, vol. 7, no. 4, pp. 1735–1738, 2016.

[5] M. Secchiero, "*FakeBook : Detecting Fake Profiles in On-line Social Networks*," *IEEE*, 2012.

[6] A. El Azab, A. M. Idrees, M. A. Mahmoud, and H. Hefny, "*Fake Account Detection in Twitter Based on Minimum Weighted Feature set*," *IEEE*, vol. 10, no. 1, pp. 13–18, 2016.

[7] M. S. B. Maind, "*Research Paper on Basic of Artificial Neural Network*," *IJRITCC*, vol. 2, no. January, pp. 96–100, 2014.

[8] V. Jakkula, "*Tutorial on Support Vector Machine (SVM)*."

[9] J. Melton, S. Buxton, H. Samet, T. J. Teorey, S. S. Lightstone, T. P. Nadeau, J. Celko, G. Ralf, M. Schneider, J. Celko, E. Cox, T. Halpin, K. Evans, P. Hallock, B. Maclean, J. Melton, J. Melton, A. R. Simon, and M. Chisholm, *Data Mining : Concepts and Techniques*. 1999.

[10] C. Science and S. Engineering, "*Research on Data Mining Classification*," vol. 4, no. 4, pp. 329–332, 2014.

[11] M. T. Hagan and M. H. Beale, "*Neural Network Design*."

[12] R. V Belavkin, "Lecture 11 : *Feed-Forward Neural Networks*," pp. 1–12.

[13] R. Sathya and A. Abraham, "*Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification*," *IJARI*, vol. 2, no. 2, pp. 34–38, 2013.

[14] I. Journal and I. Computing, "*Back-Propagation Learning Algorihtm*," *IJICIC*, vol. 7, no. 10, pp. 5839–5850, 2011.

[15] D. Sharma, "*A Study of Data Mining Clustering Techniques*," vol. 4, no. 3, pp. 490–494, 2014.

[16] H. Park and C. Jun, "*A simple and fast algorithm for K-medoids clustering," ESWA*, vol. 36, no. 2, pp. 3336–3341, 2009.

[17] S. Shah and M. Singh, "*Comparison of A Time Efficient Modified K-mean Algorithm with K-Mean and K-Medoid algorithm*," *IEEE*, pp. 0–2, 2012.

[18] C. Science, C. Application, and N. Data, "*An efficient method to improve the clustering performance for high dimensional data by Principal Component Analysis and modified*," *IJDMS*, vol. 3, no. 1, pp. 196–205, 2011.

[19] A. Saberi, M. Vahidi, and B. M. Bidgoli, "*Learn To Detect Phishing Scams Using Learning and Ensemble Methods*," *IEEE*, pp. 311–314, 2007.

[20] D. K. Srivastava and L. Bhambhu, "*Data Classification Using Support Vector Machine*," *JATIT*, 2009.

[21] G. Magno and T. Rodrigues, "*Detecting Spammers on Twitter*," *CEAS*, 2010.

[22] J. Ratkiewicz, M. D. Conover, M. Meiss, B. Gonc, A. Flammini, and F. Menczer, "*Detecting and Tracking Political Abuse in Social Media*," *AAAI*, pp. 297–304, 2010.

[23] S. Kiruthiga, "*Detecting Cloning Attack in Social Networks Using Classification and Clustering Techniques*," *IEEE*, 2014.

[24] M. Alsaleh and A. Alarifi, "*TSD : Detecting Sybil Accounts in Twitter*," *IEEE*, 2014.

[25] Y. Shen, J. Yu, K. Dong, and K. Nan, " C*hinese Micro-blogging System*," *Springer*, pp. 596–607, 2014.

[26] M. Fire, D. Kagan, A. Elyashar, and Y. Elovici, "*Friend or Foe ? Fake Profile Identification in Online Social Networks*," *IEEE*, pp. 1–27, 2012.

[27] G. Gandhi and R. Srivastava, "*Anaysis And Implementation of Modified K-Mediod Algorithm to Increase Scalability And Efficiency For Large Dataset*," *IJRET*, vol. 03, no. 06, pp. 2319–2322, 2014.

[28] E. P. Kumar and E. P. Sharma, "*Artificial Neural Networks-A Study*," *IJEERT*, vol. 2, no. 2, pp. 143–148, 2014.

[29] M. Egele, G. Stringhini, G. Stringhini, and G. Vigna, "*Towards Detecting Compromised Accounts on Social Networks*," *IEEE*, vol. 5971, no. c, 2015.

[30] D. M. Freeman and T. Hwa, "*Detecting Clusters of Fake Accounts in Online Social Networks Categories and Subject Descriptors*," *AISec*, 2015.

[31] B. Hudson, J. Matthews, S. Gurajala, J. S. White, B. Hudson, and J. N. Matthews, "*Fake Twitter accounts : Profile characteristics obtained using an activity-based*

pattern detection approach Fake Twitter accounts : Profile characteristics obtained using an activity-based pattern detection approach," *ACM*, no. August, 2015.

[32] Y. Boshmaf and K. Beznosov, "*Thwarting Fake OSN Accounts by Predicting their Victims."*

[33] S. Rahman, T. Huang, H. V Madhyastha, and M. Faloutsos, "*Detecting Malicious Facebook Applications*," *IEEE/ACM*, pp. 1–15, 2015.

[34] K. B. Kansara, "*Security against sybil attack in social network*," *ICICES*, no. Icices, 2016.

[35] A. M. Meligy, "*Identity Verification Mechanism for Detecting Fake Profiles in Online Social Networks*," *IJCNIS*, no. January, pp. 31–39, 2017.