Efficient decision making of data by using Deep learning via generating Chernoff Face

Author

KASHIF

Reg. Number

00000273683

Supervisor

Dr. MUHAMMAD USMAN AKRAM

DEPARTMENT OF COMPUTER & SOFTWARE ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY

ISLAMABAD

Jan 2021

Efficient decision making of data by using Deep learning via generating Chernoff Face

Author

KASHIF

Reg. Number

00000273683

A thesis submitted in partial fulfillment of the requirements for the degree of

MS Software Engineering

Thesis Supervisor:

Dr. MUHAMMAD USMAN AKRAM

Thesis Supervisor's Signature:

_____

DEPARTMENT OF COMPUTER & SOFTWARE ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY,

ISLAMABAD

Jan 2021

DECLARATION

I certify that this research work titled “Better decision making of data by using Chernoff faces and CNN Model*”* is my work. The work has not been presented elsewhere for assessment. The material that has been used from other sources has been properly acknowledged/referred to.

Signature of Student
Kashif
MS – 18 – CSE

PLAGIARISM REPORT (TURNITIN REPORT)

This thesis has been checked for Plagiarism. Turnitin report endorsed by Supervisor is attached.

Signature of Student

Kashif

00000273683

Signature of Supervisor

COPYRIGHT STATEMENT

ACKNOWLEDGEMENT

I am thankful to my Creator Allah Subhana-Watala to have guided me throughout this work at every step and for every new thought indeed. Whosoever helped me throughout my thesis, whether my parents or any other individual was Your will, so indeed none be worthy of praise but You.

I would also like to express special thanks to my supervisor Dr. Muhammad Usman Akram for his help throughout my thesis and GEC members Dr. Ahsan Shehzad and Dr. Arslan Shaukat for their guidance. Finally, I would like to express my gratitude to all the individuals who have rendered valuable assistance to my thesis.

ABSTRACT

In the current era in which we are living data is very important in every field of life. Data allows associations to quantify the adequacy of a given technique, When systems are instituted to conquer a test, collecting data will allow you to decide how well your good answer is performing, and your methodology should be changed or changed over the long haul. Therefore, for a better understanding and decision, we need well-arranged data. However, unfortunately, we do not get it initially. For this purpose, we need to visualize the data in such a way that we can classify between different classes. Until we do not properly classify data, we will not be able to make a good decision based on data. We have different visualization (like scatter plot, histogram, Chernoff faces, Pixel-oriented visualization, Geometric Projection Visualization, Icon-Based Visualization, Hierarchical Visualization, Visualizing Complex Data, and Relations, etc.) and classification (SVM, Decision tree, random forest, Naive Bayes, etc.) approaches for numeric data[1]. We introduce a novel approach for numeric data to increases the efficiency of the data to do a better understanding. Our approach is a combination of the visualization technique Chernoff faces[2] and the CNN model, which is used for the classification of images. For the validation of our approach, we use critical medical data hepatitis C Virus (HCV). This is very important according to the patient perspective. HCV is growing very fast worldwide. And it is the major global cause of death. It is not very dangerous in the early stages but can be deadly later on. The cause of HCV differs from country to country. In Pakistan, the key factor of HCV is the reuse of glass syringes. In Europe and America, the cause of HCV is the use of unsafe drug injections. Classification in this type of dataset is very important because it can save a life. Without using our approach, the efficiency of the data was about 22% to 26% and after using our proposed approaches the efficiency of data to increase up to 99%.

Key Words: Data classification, Data efficiency, Feature extraction using the CNN model, Data visualization using Chernoff faces

TABLE OF CONTENT

DECLARATION iii

PLAGIARISM REPORT (TURNITIN REPORT) iv

COPYRIGHT STATEMENT v

ACKNOWLEDGEMENT vi

LIST OF FIGURES xi

LIST OF TABLES xi

CHAPTER 1. INTRODUCTION 1

1.1. MOTIVATION 1

1.2. PROBLEM STATEMENT 2

1.3. AIMS AND OBJECTIVES 2

1.4. STRUCTURE OF THESIS 3

CHAPTER 2. DECISION MAKING, DATA MINING, AND CHERNOFF FACES 4

2.1. DECISION MAKING 4

2.1.1. DECISION MAKING BASED UPON DATA 4

2.1.2. WHY IT IS IMPORTANT 5

2.2. DATA VISUALIZATION 6

2.2.1. ICON-BASED VISUALIZATION TECHNIQUE 6

2.2.2. DATA MINING 8

2.2.3. HOW DATA-MINING WORKS 9

2.2.4. DATA WAREHOUSING AND MINING SOFTWARE 9

2.2.5. DATA MINING AND MACHINE LEARNING 10

2.2.6. DATA USE 10

2.2.7. FOUNDATION FOR LEARNING 11

2.2.8. PATTERN RECOGNIZATION 11

2.2.9. IMPROVED ACCURACY 12
2.3. CLASSIFICATION 13
2.3.1. WHAT IS CLASSIFICATION? 14
2.3.2. WHAT IS PREDICTION? 14
2.3.3. HOW DOES CLASSIFICATION WORKS? 15
CHAPTER 3. LITERATURE REVIEW 18
3.1. CHERNOFF FACES 18
3.2. CLASSIFICATION 20
3.3. CONVOLUTION NEURAL NETWORK (CNN) 25
CHAPTER 4. METHODOLOGY 28
4.1. DATASET COLLECTION 29
4.2. PRE-PROCESSING 30
4.3. FEATURE SELECTION 31
4.3.1. FILTER METHOD 32
4.3.2. WRAPPER METHOD 33
4.3.3. CORRELATION WITH HEATMAP 33
4.4. MAKING CHERNOFF FACES 34
4.4.1. FEATURE DESCRIPTION 36
4.5. FEATURE EXTRACTION USING CNN MODEL 38
4.5.1. LIBRARIES 39
4.5.2. CONVOLUTION NEURAL NETWORK 40
4.5.3. RESEMBLANCE WITH MLP 40
4.6. CLASSIFICATION 43
4.6.1. DECISION TREE 44
4.6.2. RANDOM FOREST 45

4.6.3. SVM 45
4.6.4. BAGGING 47
4.7. SUMMARY 47
CHAPTER 5. RESULTS AND DISCUSSION 48
5.1. DATASET 48
5.2. DATASET DIVISION 48
5.3. PRE-PROCESSING AND FEATURE SELECTION 49
5.3.1. UNIVARIATE SELECTION 50
5.3.2. FEATURE IMPORTANCE 51
5.3.3. CORRELATION WITH HEATMAP 52
5.4. CHERNOFF FACES 54
5.4.1. ATTRIBUTES ASSIGNMENT 54
5.4.2. MAKING CHERNOFF FACES 55
5.4.3. IMAGE SLICING 56
5.5. FEATURE EXTRACTION USING CNN 57
5.6. CLASSIFICATION RESULTS 59
5.6.1. ORIGINAL DATASET RESULTS 59
5.6.2. SELECTED ATTRIBUTES DATASET RESULTS 61
5.6.3. EXTRACTED FEATURES DATASET RESUL 64
CHAPTER 6. CONCLUSION & FUTURE WORKs 69
6.1. CONCLUSION 69
6.2. CONTRIBUTION 69
6.3. FUTURE WORK 70
REFERENCES 71

LIST OF FIGURES

Figure 2.1: Chernoff Faces[2] 7
Figure 2.2: Sticky Figures[6] 8
Figure 2.3: Building Classifier Model[9] 16
Figure 2.4:Classifier For The Classification[10] 16
Figure 4.1: Flow of the Proposed Methodology 29
Figure 4.2: World Wide Effects of Hepatitis C Virus[37] 30
Figure 4.3: Feature Selection Techniques 32
Figure 4.4: Filter Method for Feature Selection 32
Figure 4.5: Wrapper Method for Feature Selection 33
Figure 4.6: Chernoff Faces[2] 35
Figure 4.7: Bad Assignment of Features for Making Chernoff of Faces[38] 36
Figure 4.8: Good Feature Assignment for Chernoff Faces[39] 37
Figure 4.9: CNN Model for The Features Extraction 39
Figure 4.10: Structure of the CNN Model 41
Figure 4.11: Decision Tree Classifier[40] 44
Figure 4.12: Random Forest Classifier[41] 45
Figure 4.13: SVM Different Hyper Plans 46
Figure 5.1: Dataset Division Code 49
Figure 5.2: Best 17 Features Extracted Using Feature Importance 52
Figure 5.3: Priority Matrix of n*n Feature for Having 17 High Priority Features 53
Figure 5.4: Chernoff Faces of All Records of One Output Label 56
Figure 5.5: Image Slicing to Get Number of Images Equal to the Number of Records 57
Figure 5.6: Detailed Out Put of the CNN Model 58
Figure 5.7: Our Proposed Methodology Results 68
Figure 5.8: Results of past Researches 68

LIST OF TABLES

Table 3.1: Literature Review of Chernoff Faces 20

Table 3.2: Literature Review of Supervised data classifiers 24
Table 3.3: A literature review of CNN Classifier with numeric classifier 27
Table 4.1: Dataset after Pre-Processing 31
Table 4.2: Coefficient for Different Datasets 32
Table 5.1: Dataset after Pre-Processing 50
Table 5.2: Best 17 Features Using Univariate Selection Technique 51
Table 5.3: Result of CNN Model 58
Table 5.4: Accuracy results of four classifiers on Original dataset 59
Table 5.5: Confusion Matrix of Decision Tree 59
Table 5.6: Confusion Matrix of Random Forest 60
Table 5.7: Confusion Matrix of SVM 60
Table 5.8: Confusion Matrix of Bagging 61
Table 5.9: Accuracy score of four classifiers using selected attributes dataset 61
Table 5.10: Confusion Matrix of Decision Tree 62
Table 5.11: Confusion Matrix of Random Forest 62
Table 5.12: Confusion Matrix of SVM 63
Table 5.13: Confusion Matrix of Bagging 63
Table 5.14: Accuracy results of four classifier using extracted features 64
Table 5.15: Confusion Matrix of Decision Tree 64
Table 5.16: Confusion Matrix of Random Forest 65
Table 5.17: Confusion Matrix of SVM 65
Table 5.18: Confusion Matrix of Bagging 66
Table 5.19: Results for All Three Dataset Using Different Approaches 66
Table 5.20: Past Results using Python and R language 67

CHAPTER 1. INTRODUCTION

In the previous years, a lot of research has been done on the deep learning model and a few on Chernoff's face. Deep learning comes with the revolution in the field of decision-making and data analysis. As we know, data is very important in every field and decision-making depends on the previous data. However, an un-arranged form of data is nothing but a waste for everyone. So to use the data in such a way that we can understand it and make the decision on its basis we need to arrange the data in a way we can clearly understand the flow and the formation of data. Initially, data is in a different form, full of anomalies, and blunders. This is a challenge itself to visualize the data in a uniform state. For this purpose, we need to apply preprocessing to the data. Which have some sub-processes like data acquisition; Import all the crucial libraries, import the dataset, identifying and handling the missing values, Encoding the categorical data, Splitting the dataset, and Feature scaling. After preprocessing we are unable to do perform different processes on the data for the decision-making. Data preprocessing, data visualization, and data classification techniques are used for increasing the efficiency of the data so we can make better decisions. A lot of supervised classification techniques for data like (Random forest, Naive Bayes, Bagging, SVM, Decision tree, etc.) are used but sometimes data is very critical so we cannot classify it properly to its maximum limits and until we are not able to classify the data we cannot able to make a good decision. Our study aims to properly visualize and classify the data for those people's doctors, engineer, and Business planner) who are eligible to decide based on data. And the decision they will make will be very crucial for those peoples (Patient, Business owner) who will be affected by the decision. As much as correctly we will classify the data it will be beneficial for the stakeholders who are associated with that data.

1.1. MOTIVATION

CNN models are used for the classification of the data which is in image form and also works with numeric or supervised data but is difficult to deal with the numeric data when we have a large number of attributes but for images, it's easy to use the CNN models. CNN is very efficient in the classification. On the other hand different supervised classifiers are used for the classification and also efficient as well but not like the CNN model. Our study aimed to

introduce a novel approach that can help us to classify our supervised data with the CNN model without any difficulty. So we introduced a novel approach which is a combination of Chernoff faces and the CNN model. This combination has a logical purpose as CNN efficiently works with the images so we use the Chernoff visualization technique which visualizes the data in form of human-like faces as images. And these human-like face images will be the input for the CNN model. CNN model will use these images and will extract the important features from them. Furthermore, some supervised classifier will be used for the classification of those features which is extracted through CNN. This is very efficient as well. We proved the effectiveness of the proposed methodology using a critical medical dataset which is known as Hepatitis C Virus (HCV).

1.2. PROBLEM STATEMENT

In the current era, we use a lot of data to make a decision. For better decisions, we need better data. Little change in the data can create a major effect on the result. For an understanding of data, we visualize the data in a form in which we classify the data into different classes. So we need an approach which can remove the little chance of mistake. We discuss the medical data which is very important because it belongs to human health. In such cases, little mistakes can be dangerous for someone's life. We introduce a novel approach that is very useful in such cases. For validation of our approach, we use the HCV dataset. This is very complex and has less accuracy while using classifiers used for supervised learning. So we chose this dataset for the validation of our approach to the effectiveness of the methodology can be seen clearly. Max's accuracy achieved using this dataset using a multiclass label classifier was about 26% using python and max accuracy using binary label classifier was about 51%. But we excellently prove our study best by using the multiclass label classifier. And we increase the efficiency of the dataset by up to 99%. This is a very high accuracy to be achieved.

1.3. AIMS AND OBJECTIVES

The major objectives of the research are as follow:

- To increase the efficiency of data
- To do the better classification of data
- Very efficient very the dimensionality is very high

- Efficient use of the memory
- Provide better results in similar data records
- To set a milestone for future applications
- Decision making in Robots

1.4. STRUCTURE OF THESIS

This structure of the thesis is as follows:

Chapter 2: covers the importance of decision making and visualization techniques. It further discusses deep learning and some supervised classifier for the classification.

Chapter 3: a review of the literature and the significant work done by researchers in the past few years for the Chernoff faces supervised classifier and CNN model for feature extraction.

Chapter 4: consists of the proposed methodology in detail. It includes three main modules: visualization of data, feature extraction using the CNN model followed by their classification

Chapter 5: introduces the databases used for evaluation purposes. All the experimental results are discussed in detail with all desired figures and tables.

Chapter 6: concludes the thesis and reveals the future scope of this research

CHAPTER 2. DECISION MAKING, DATA MINING, AND CHERNOFF FACES

2.1. DECISION MAKING

The term 'Big data' alone has become something of a popular expression lately and in light of current circumstances[3][4][2]. By using the bounty of automated encounters open promptly accessible and getting a handle on the force of business understanding, it's possible to make more instructed decisions that will incite business improvement, headway, and an extended essential concern. By executing the right noteworthy contraptions and perceiving how to look at similarly as to evaluate your data unequivocally, you will have the choice to choose such a data driven decisions that will drive your business forward. This sounds fabulous on a fundamental level. However, basically, whether or not you approach the world's most critical data, it's possible to make decisions that excuse generous agreement, going with your gut in light of everything. Generally speaking, this can exhibit negative to the business. While here and there it's alright to follow your impulses, by far most of your business-based choices ought to be upheld by measurements, realities, or figures identified with your points, objectives, or activities that can guarantee a steady spine to your administration reports and business tasks. To help you on your journey towards scientific illumination, we will investigate data-driven dynamic, study the significance of data-driven dynamic, and inspect some certifiable instances of transforming knowledge into business-boosting activity.

2.1.1. DECISION MAKING BASED UPON DATA

Decision making because of data is a cycle that remembers gathering data subordinate for quantifiable targets or KPIs, separating models and real factors from these pieces of data, and utilizing them to make frameworks and activities that advantage the business in different regions. On an extremely fundamental level, data-driven powerful techniques seeking after key business targets by using checked examined data rather than just shooting in indefinite quality. In any case, to remove genuine motivating force from your data, it must be precise similarly as appropriate to your focuses. A social affair, removing, coordinating, and looking at pieces of data

for overhauled data-driven elements in business was before a far-reaching task, which regularly conceded the entire data dynamic cycle. However, today, the new development and democratization of business information programming draw in customers without significant joined specific capacity to separate similarly as concentrate encounters from their data. As a quick result, less IT maintain is expected to make reports, examples, recognitions, and pieces of information that support the data dynamic cycle. From these unforeseen developments, data science was imagined (or perhaps, it progressed enormously) – a request where hacking aptitudes and experiences meet forte capacity. This new calling incorporates separating a ton of unrefined data to choose wise data that is driven by business decisions. The 'gold' that data specialists 'mine' comes in two specific sorts: emotional and quantitative, and both are essential to making a data-driven decision. Abstract assessment bases on data that isn't described by numbers or estimations, for instance, gatherings, chronicles, and records. Abstract data assessment relies upon observation rather than assessment. Here, it's crucial to coding the data to ensure that things are gathered methodically similarly as distinctly. Quantitative data examination bases on numbers and bits of knowledge. The center, standard deviation, and other illustrative subtleties expect a dire part here. This sort of assessment is assessed rather than viewed. Both emotional and quantitative data should be researched to choose more shrewd data-driven business decisions. Since we've explored the hugeness of dynamic in business, it's an ideal occasion to consider the inspiration driving why Decision making dependent on is critical. "Information is the oil of the 21st Century, and the examination is the start engine." Peter Sondergaard

2.1.2. WHY IT IS IMPORTANT

The centrality of data in decision lies in consistency and steady turn of events. It enables associations to make new business openings, make more pay, anticipate future examples, improve current operational undertakings, and produce essential pieces of information. That way, you stay to create and propel your space after some time, making your affiliation more flexible, in this manner. The serious world is in a predictable state of progress, and to move with the ever-changing scene around you, you should utilize information to make more instructed and momentous information-driven business decisions. Information driven business decisions speak to the critical point in time for associations. This is a show of the criticalness of online data

recognition in unique. MIT Sloan School of Management teachers Andrew McAfee and Erik Brynjolfsson once explained in a Wall Street Journal article that they played out an examination identified with the MIT Center for Digital Business. In this examination, they found that among the associations surveyed, the ones that were chiefly data-driven benefitted by 4% higher productivity similarly as 6% higher advantages. Associations that approach dynamic helpfully will all in all view information as a certified asset more than associations with other, more questionable strategies.

2.2. DATA VISUALIZATION

The perception of huge multivariate data indexes is as yet a difficult errand, particularly when we think about the investigation of an assortment of traits in a single portrayal. Numerous methods have been created to imagine such multivariate data collections. Different visualization methods are the following

- Pixel-oriented visualization techniques
- Geometric projection visualization techniques
- Visualizing complex data and relations
- Icon-based visualization techniques
- Hierarchical visualization techniques

But our main concern is about discussing the icon-based visualization technique, which is also part of our proposed methodology.

2.2.1. ICON-BASED VISUALIZATION TECHNIQUE

Visualization of the data values as features of icons
Typical visualization methods

- Chernoff Faces
- Stick Figures

General techniques

- Shape coding: Use the shape to represent certain information encoding
- Color icons: Use color icons to encode more information

- Tile bars: Use small icons to represent the relevant feature vectors in document retrieval

2.2.1.1. CHERNOFF FACES

Chernoff's faces were introduced in 1973 by statistician Herman Chernoff[5]. Show multidimensional information of up to 18 factors (or measurements) as an animation human face. Utilizing Asymmetrical appearances, upto36 measurements can be shown. An approach to show factors on a two-dimensional surface, e.g., let x be eyebrow incline, y be eye size, z be nose length, and so forth The figure shows faces created utilizing 10 qualities - head unconventionality, eye size, eye dispersing, eye flightiness, understudy size, eyebrow incline, nose size, mouth shape, mouth size, and mouth opening): Each allocated one of 10 potential qualities, generated using *Mathematica*(S. Dickson)

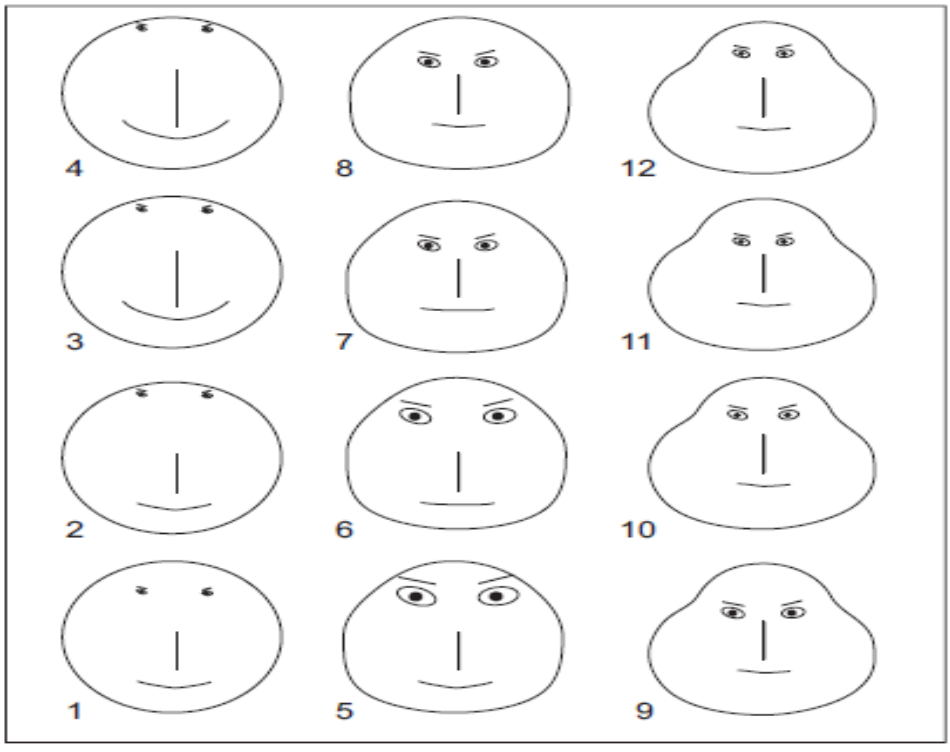

Figure 2.1: Chernoff Faces[2]

2.2.1.2. STICKY FIGURES

A stick figure is a basic drawing of an individual or creature[6], made out of a couple of lines, bends, and specks. In a stick figure, the head is spoken to by a circle, once in a while adorned with subtleties, for example, eyes, mouth, or hair. The arms, legs, and middle are generally spoken to by straight lines. Subtleties, for example, hands, feet, and a neck might be available or missing, and the more straightforward stick figures frequently show a vague passionate articulation or unbalanced appendages.

Spray painting of stick figures is found since the beginning, regularly scratched with a sharp item on hard surfaces, for example, stone or solid dividers. Stick figures are frequently utilized in portrays for film storyboarding.

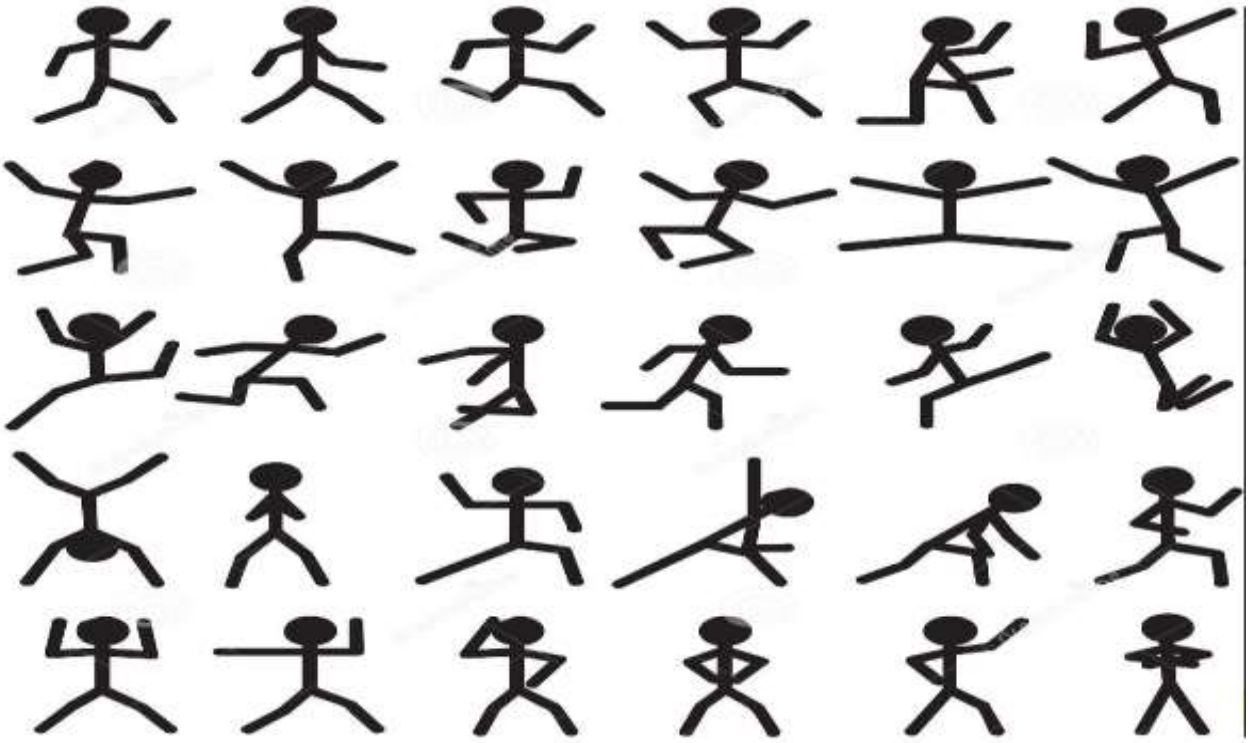

Figure 2.2: Sticky Figures[6]

2.2.2. DATA MINING

Data mining is a cycle used by associations to change rough data into supportive information. By using programming to look for plans in tremendous bunches of data, associations can consider their customers to develop additionally convincing exhibiting systems, increase bargains, and decreasing costs. Data mining depends on incredible data combination, warehousing, and PC

planning. Data mining measures are utilized to assemble AI models that power applications including web crawler innovation and site proposal programs.

2.2.3. HOW DATA-MINING WORKS

Data mining incorporates researching and analyzing colossal squares of information to assemble critical models and examples. It will in general be used in a combination of ways, for instance, data-based advancing, credit peril the board, distortion acknowledgment, spam Email isolating, or even to perceive the thought or appraisal of customers.

The Data mining measure isolates into five phases. Most importantly, affiliations accumulate data and weight it into their data stockrooms. Next, they store and manage the data, either on in-house laborers or the cloud. Business specialists, administrative gatherings, and information advancement specialists access the data and choose how they have to assemble it. By then, application programming sorts the data reliant on the customer's results, ultimately, the end-customer presents the data in an easy-to-share plan, for instance, a graph or table.

2.2.4. DATA WAREHOUSING AND MINING SOFTWARE

Data mining programs examine connections and examples in information dependent on what clients demand. For instance, an organization can utilize information mining programming to make classes of data. To delineate, envision a café needs to utilize information mining to decide when it should offer certain specials. It takes a gander at the data it has gathered and makes classes dependent on when clients visit and what they request. In different cases, information diggers discover groups of data dependent on legitimate connections or take a gander at affiliations and successive examples to finish up patterns in buyer conduct. Warehousing is a significant part of information mining. Warehousing is when organizations incorporate their information into one data set or program. With an information distribution center, an association may turn off portions of the information for explicit clients to investigate and utilize. Notwithstanding, in different cases, examiners may begin with the information they need and make an information stockroom dependent on those specs. Despite how organizations and different elements sort out their information, they use it to help the board's dynamic processes.

2.2.5. DATA MINING AND MACHINE LEARNING

With large information getting so common in the business world[7][8], a lot of information terms will in general be tossed around, with numerous not exactly understanding what they mean. What is information mining? Is there a contrast between AI versus information science? How would they associate with one another? Isn't AI simply man-made consciousness? These are acceptable inquiries, and finding their answers can give a more profound, all the more compensating comprehension of information science and investigation and how they can profit an organization.

2.2.6. DATA USE

Both information mining and AI are established in information science and by and large fall under that umbrella. They regularly meet or are mistaken for one another, yet there is a couple of key differentiation between the two. Here's a gander at some information mining and AI contrasts between information mining and AI and how they can be utilized.

One key contrast between machine learning and data mining is how they are utilized and applied in our regular day to day existences. For instance, information mining is regularly utilized by AI to see the associations between connections. Uber utilizes AI to compute ETAs for rides or feast conveyance times for UberEATS.

Data mining can be utilized for an assortment of purposes, including money related examination. Speculators may utilize information mining and web scratching to take a gander at a beginning up's financials and help decide whether they need to offer to subsidize. An organization may likewise utilize data mining to help gather information on deal patterns to more readily educate all from promoting to stock requires, just as to make sure about new leads. Information mining can be utilized to go over web-based media profiles, sites, and advanced resources for arranging data on an organization's optimal prompts to start an effort crusade. Utilizing data mining can prompt 10,000 leads in a short time. With this much data, and information the researcher can even foresee future patterns that will enable an organization to get ready well for what clients may need in the months and years to come. Machine-learning exemplifies the standards of information mining, however can likewise make programmed relationships and gain from them to apply to new calculations. It's simply the innovation driving vehicles that can rapidly change following new conditions while driving. AI likewise gives moment suggestions when a purchaser buys an item from Amazon. These calculations and examinations are continually

intended to be improving, so the outcome will just get more exact after some time. AI isn't man-made reasoning; however, the capacity to learn and improve is as yet a great accomplishment. Banks are now utilizing and putting resources into AI to help search for extortion when charge cards are swiped by a merchant. Citibank put resources into worldwide information science venture Feedzai to distinguish and destroy monetary extortion continuously across on the web and in-person banking exchanges. The innovation serves to quickly distinguish extortion and can assist retailers with securing their monetary movement.

2.2.7. FOUNDATION FOR LEARNING

Both data mining and machine learning draw from a similar establishment, yet in various ways. An information researcher utilizes information mining pulls from existing data to search for rising examples that can help shape our dynamic cycles. The garments brand Free People, for instance, utilizes data mining to search over great many client records to shape their search for the season. The information investigates top of the line things, what was restored the most, and client input to help sell more garments and improve item suggestions. This use of data assessment can incite an improved customer experience as a rule.

Machine learning, on the other hand, can pick up from the current data and give the foundation imperative to a machine to teach itself. Zebra Medical Vision developed a machine learning estimation to anticipate cardiovascular conditions and capacities that lead to the downfall of more than 500,000 Americans consistently. AI can look at plans and addition from them to change lead for future events, while data mining is conventionally used as an information hotspot for machine learning to pull from. Even though data specialists can set up information digging to therefore looking for unequivocal kinds of data and limits, it doesn't learn and apply data in isolation without human cooperation. Data mining moreover can't normally notice the association between existing pieces of data with a comparable significance that machine learning can.

2.2.8. PATTERN RECOGNIZATION

Gathering data is only fundamental for the test; the other part is sorting out everything. The right programming and devices are ought to have been prepared to explore and interpret the goliath proportions of data analysts accumulate and find obvious guides to catch up on. Something

different, the data would commonly be unusable aside from if data analysts could submit their opportunity to look for these complexes, as often as possible subtle and self-assertive models isolated. Additionally, anyone even somewhat familiar with data science and data assessment understands this would be a debilitating, dreary task.

Associations could use data to shape their business envisioning or sort out what sorts of things their customers need to buy. For example, Wal-Mart accumulates the reason for bargains from over 3,000 stores for its data circulation focus. Shippers can see this information and use it to recognize buying practices and guide their stock desires and cycles for what's to come. The realities affirm that data mining can reveal a couple of models through courses of action and gathering assessment. Regardless, AI makes this thought a step further by using comparative estimations data mining uses to therefore pick up from and conform to the assembled data. As malware transforms into an irrefutably inevitable issue, AI can look for plans in how data in structures or the cloud is gotten to. Computer-based intelligence similarly observes guides to help perceive which records are malware, with a raised degree of accuracy. This is overseen without the prerequisite for predictable perception by a human. In case unpredictable models are perceived, an alert can be passed on so a move can be made to prevent the malware from spreading.

2.2.9. IMPROVED ACCURACY

Both data mining Furthermore, machine learning can help improve the exactness of information accumulated. Regardless, information mining and how it's dismembered overall identify with how the data is composed and accumulated. Information mining may join using eliminating and scratching programming to pull from a large number of resources and channel through data that investigators, data scientists, money related masters, and associations use to look for models and associations that help improve their primary concern. One of the fundamental foundations of machine learning is information mining. Data mining can be used to eliminate more careful data. This finally refines your machine learning to achieve better results. An individual may miss the various affiliations and associations between data, while machine learning advancement can pinpoint these moving pieces to arrive at an outstandingly exact assurance to help shape a machine's direct. AI can improve relationship information in CRM structures to help bargain

bunches better grasp their customers and make a relationship with them. Gotten together with AI, an association's CRM can analyze past exercises that lead to a change of buyer analysis. It can similarly be used to sort out some way to predict which things and organizations will offer the best and how to shape elevating messages to those clients.

2.3. CLASSIFICATION

Classification is a data mining limit that gives out things in a combination to target characterizations or classes. The goal of characterization is to decisively predict the target class for each case in the information. For example, a gathering model could be used to perceive progressed competitors as low, medium, or high credit possibilities. A gathering task begins with an enlightening record in which the class errands are known. For example, an arrangement model that predicts credit peril could be made reliant on watched information for some serious up-and-comers all through some unclear time. Despite the recorded FICO score, the information may follow work history, a home belonging or rental, extended lengths of home, number and sort of theories, and so on FICO evaluation would be the target, various credits would be the pointers, and the information for each customer would contain a case.

Groupings are discrete and don't recommend demands. Constant, floating-point regards would show a numerical, instead of an outright, target. An insightful model with a numerical target utilizes a backslide figuring, not a gathering estimation. The simplest sort of grouping issue is a combined request. In the matched course of action, the target attribute has only two likely characteristics: for example, high FICO evaluation or low FICO score. Multiclass targets have various characteristics: for example, low, medium, high, or dark FICO appraisal. In the model structure (planning) measure, a course of action figuring finds associations between the assessments of the markers and the assessments of the goal. Unmistakable request figurings use different techniques for finding associations. These associations are summarized in a model, which would then have the option to be applied to another information assortment where the class assignments are dark. Request models are attempted by standing out the foreseen characteristics from acknowledged target characteristics in a lot of test information. The valid information for a grouping adventure is normally isolated into two information records: one for building the model; the other for testing the model. See "Testing a Classification Model". Scoring a gathering model results in-class assignments and probabilities for each case. For

example, a model that portrays customers as low, medium, or high worth would in like manner foresee the probability of each request for each customer. The order has various applications in customer division, business showing, promoting, the credit assessment, and biomedical and drug response demonstrating.

Two types of information investigation can be utilized for removing models depicting significant classes or to anticipate future information patterns. These two structures are as per the following

- Classification
- Prediction

Classification models foresee clear cut class names, and forecast models anticipate constant esteemed capacities. For instance, we can fabricate a characterization model to arrange bank advance applications as either protected or hazardous or an expectation model to foresee the consumptions in dollars of likely clients on PC hardware given their pay and occupation.

2.3.1. WHAT IS CLASSIFICATION?

Following are the instances of situations where the information examination task is Classification

- A bank advance official needs to investigate the information to know which client (credit candidate) are hazardous or which are protected.

- A showcasing director at an organization needs to break down a client with a given profile, who will purchase another PC.

In both of the above models, a model or classifier is built to foresee the downright marks. These names are dangerous or ok for credit application information and yes or no for showcasing data.

2.3.2. WHAT IS PREDICTION?

The following are instances of situations where the information examination task is Prediction. Assume the advertising director needs to anticipate how much a given client will spend during a deal at his organization. In this model, we are tried to foresee a numeric worth. In this manner, the information examination task is a case of numeric expectation. For this situation, a model or an indicator will be developed that predicts a constant esteemed capacity or requested worth. Relapse investigation is a factual strategy that is regularly utilized for numeric prediction.

2.3.3. HOW DOES CLASSIFICATION WORKS?

With the assistance of the bank advance application that we have examined above, let us comprehend the working of arrangement. The Data Classification measure incorporates two stages

- Building the Classifier or Model
- Using Classifier for Classification

2.3.3.1. BUILDING CLASSIFIER OR MODEL

- This step is the learning step or the learning stage.
- In this progression, the classification algorithms fabricate the classifier.
- The classifier is worked from the preparation set comprised of information base tuples and their related class names.
- Each tuple that comprises the preparation set is alluded to as a classification or class. These tuples can likewise be alluded to as a test, object, or data points.

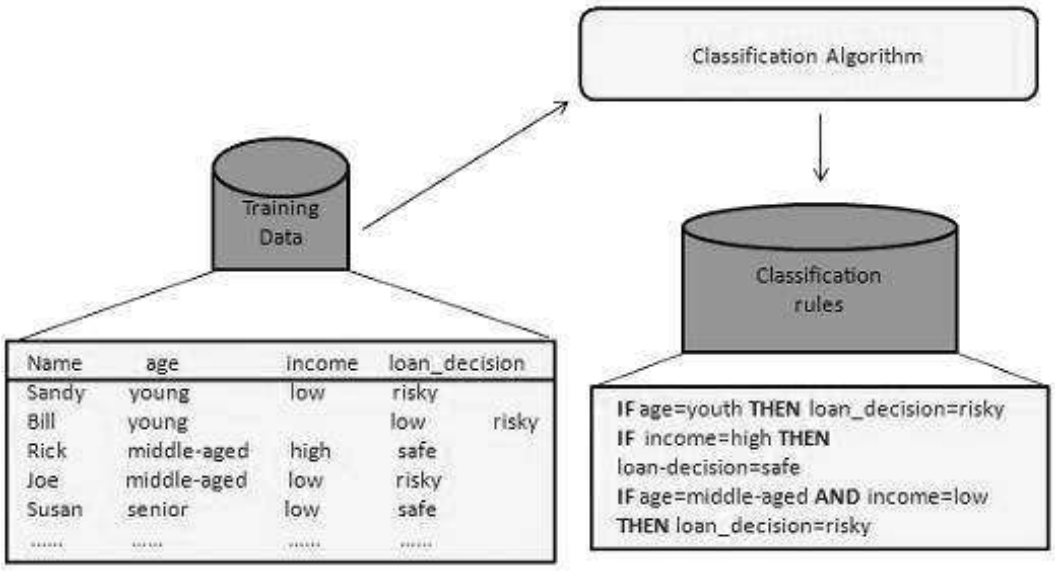

Figure 2.3: Building Classifier Model[9]

2.3.3.2. Using Classifier for Classification

In this step, the classifier is utilized for characterization. Here the test information is utilized to appraise the exactness of grouping rules. The characterization rules can be applied to the new information tuples if the exactness is thought acceptable.

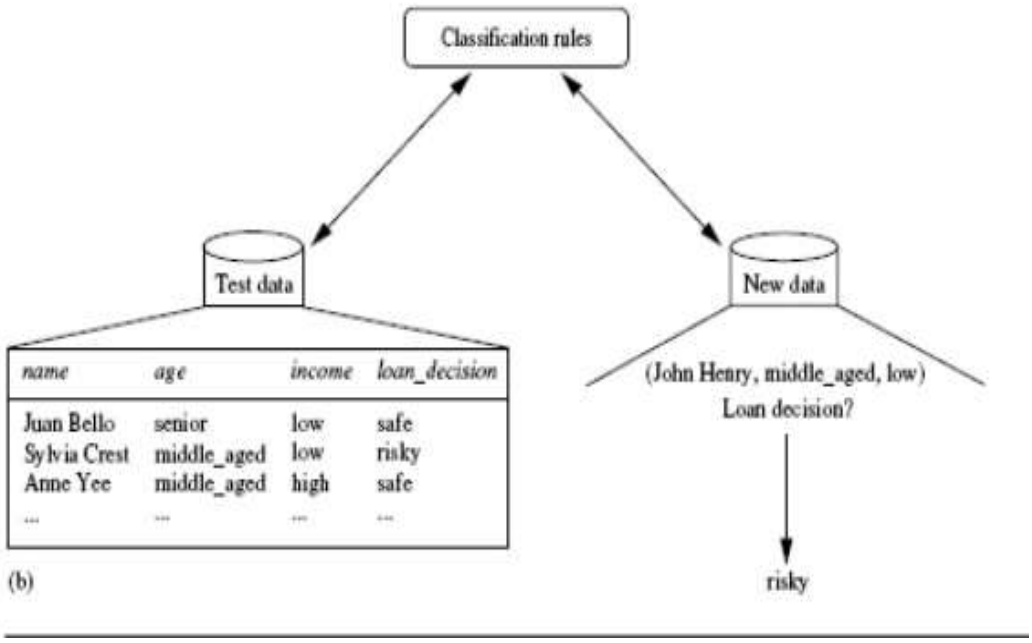

Figure 2.4:Classifier For The Classification[10]

2.3.3.3. Classification and Prediction Issues

The major issue is preparing the data for Classification and Prediction. Preparing the data involves the following activities

- Data Cleaning: Data cleaning includes eliminating the clamor and treatment of missing qualities. The commotion is taken out by applying smoothing procedures and the issue of missing qualities is understood by supplanting a missing incentive with the most generally happening an incentive for that trait.

- Relevance Analysis: The data information base may likewise have unimportant credits. Relationship examination is utilized to know whether any two given ascribes are connected.
- Data Transformation and reduction: The data can be transformed by any of the following methods.
    - Normalization: The data is changed utilizing standardization. Standardization includes scaling all qualities for an offered property to make them fall inside a little indicated range. Standardization is utilized when in the learning step, the neural organizations or the strategies including estimations are utilized.
    - Generalization: The data can likewise be changed by summing it up to the higher idea. For this reason, we can utilize the idea of hierarchies.

CHAPTER 3. LITERATURE REVIEW

Data has always been a matter of concern for every field. So the main aim of data mining or deep learning is to increase the efficiency of the data so we can do better decisions on its basis. Every dataset has its worth in its field. Several techniques and proposed methodologies are used on different types of datasets to improve their accuracy. We talk about the medical field related dataset which is very important due to its relation with human life. By increasing the efficiency of the dataset the decision-making of the viewer or doctor can be better. And this better understanding of the doctor for making a decision can save many lives. The aim of our study is also to increase the efficiency of the dataset so the decision making based on data can be better to get the benefits. The literature review is discussed below all about the different supervised and visualization techniques and on CNN models and the effectiveness of the implemented models while improving the accuracy of the dataset. As we know CNN models are very effective in the classification of the images and also during the extraction of the important features from the input images, so our purpose is to use this functionality of the CNN models for the improvement of the accuracy of the supervised numeric datasets. It will prove an important asset for the complex datasets having less accuracy while classification. And also will be very beneficial for the peoples who are related to that dataset

3.1. CHERNOFF FACES

Hongjie Zhang et al.[11] Used the Chernoff faces for the visualization of the resistant spot welding quality assessment. Obtaining high quality at a low cost is a big challenge for the automotive industry. Earlier industries used conventional destructive test methods but they bring labor cost and waste. So they use the signal acquisition device for the data collection during resistant spot welding. Different values are obtained using a commercial Rogowsky coil which is used to measure the welding current signals. And this use to calculate the value of welding current. They calculate 9 different features from it and based on these features they construct Chernoff faces having four different classes (Poor, Good, Excellent, and Expulsion). Based on these facial expressions they excellently classify the quality of the resistant spot welding.

Kim et al.[12] Use the Chernoff faces visualization technique for the comparison of library service in public libraries of two different cities of two different countries. One city is London

and the other one is Seoul which is the capital of South Korea. The study aimed to analyze the services which were determined by the different features of the library as the number of staff, number of professionals, number of collection, library budget, etc. These features were used to construct the Chernoff faces for the visualization to see the variance between the performance in the public service library in London and Seoul. They use the Chernoff technique for data visualization because it covers multi-features at a time and it's very useful while handling big data. The result of this study was London's performance was good than Seoul's. And Seoul needs more concentration on the resources for improving the public library services.

Eduard Gerhardt et al.[13] Use the Chernoff faces for the visualization of the big data. They use this technique for the annual calculation of the profit or loss. They assign different values to different features. Like Curvature of the mouth to Annual Revenue, Size of the eye to annual cost, a form of the chin to profit, the height of the face to Return on investment, Length of the nose to customer satisfaction, and length of the eyebrow to employee satisfaction. As the values of the attributes increase the size of the feature is also increased accordingly. This helps to identify at what place the company is gaining profit or at what place the company is in the loss. They tested this technique on several companies and concluded that if the difference between the companies is very small then the results could be identified which can be a problem.

Ruixia Song et al.[14] Define proposed techniques that are used to classify the Chernoff faces using the V-System. They use two examples to validate the approach. First of all the use, the Chernoff face visualization technique for the data visualization of 22 athletes, and then they use the v-system approach to classify these faces into four categories Excellent, Good, Moderate, and poor. And in the second example, they use the record of 30 companies and then produced the Chernoff faces based upon their data. And then they classify the data into the same four categories using v-system.

Salman Abdul Moiz et al.[15] Uses the Chernoff faces techniques for the visualization level or code for smelling the number of blunders of mistake that a program or project has. In the current era, lots of organizations are developing projects in the agile methods and this method needs quick feedback of the developed module so that they can move further. For the purpose to evaluate the level of anomalies they used Chernoff visualization techniques so they can say the performance of the developed module. This approach proved beneficial for both knowledge and data-driven approach. These visualization techniques detect the bad code 95% correctly.

Arto Haara et al.[16] Use the different visualization techniques for the forest data. The main purpose of the study was to do better planning for the forest decision planning problem. First of all, they introduce three levels of planning. At the first level decisions are made at the standing state thus the forest resource data are used. At the second level, they also use the stand state data but there are more possibilities to use elaborated data. At the third level, they use holding level data for the planning. They encourage many forest owners to participate in the planning process to define management proposals so they can meet their goals. After the collection of data, they use different visualization techniques to have a proper view of data. So Chernoff faces proved as one of the best technique which covers all three levels for decision planning problem.

Table 3.1: Literature Review of Chernoff Faces

Year	Author	Technique	Dataset
2015	Hongjie Zhang et al.[11]	Chernoff	Spot welding data
2017	Kim et al.[12]	Chernoff	Public library Data
2019	Eduard Gerhardt et al.[13]	Chernoff	Car sensor data
2010	Ruixia Song et al.[14]	Chernoff	Athletes data
2019	Salman Abdul Moiz et al.[15]	Chernoff	Programming code data
2018	Arto Haara et al.[16]	Chernoff	Forest decision planning data

3.2. CLASSIFICATION

Satish CR Nandipati et al.[17] Defines the comparison between the two classification approaches of the HCV dataset. They show the results of the classification of data by using a binary classifier and multiclass classifier and also use different language to justify the approach. They use Python and R as different tools or languages. They use feature selection techniques to collect

the best features for this they use priority matrix, heatmap. After this, they made three sets of features 12 features, 21 features, and 29 features. After this, they apply two different approaches using python which are binary class and multiclass. And then they did follow the same process by using R. After all implementation of these processes the result show that which results they got with R is better than the Results which is collected by python. The maximum accuracy with multiclass using python with all three different features set was 24.8% and with the binary class, the maximum accuracy was 50.39%. And the Accuracy with the R using multiclass using all three feature sets was 50.12% and the accuracy for the binary class by using three different sets of features was 50.59%. These results clearly show the R is performing better than python in multiclass and binary classes as well.

Abd El-Salam et al.[18] [19][20]In another examination, one of the inconveniences in interminable liver illness is esophageal varices. Thus is it important to know the presence of the esophageal varices through upper gastrointestinal endoscopy to abstain from bleeding. This strategy builds the remaining task at hand of endoscopy units since fluctuates are available in under half of the patients with cirrhosis, and this technique is awkward for some patients. Hence, the forecast of varices by noninvasive techniques profits by upper gastrointestinal endoscopy screening. Past examinations indicated that Fibrosis-4 record (FIB-4) can be valuable to foresee esophageal varices with 66.9 exactness and 63% Area Under the Curve (AUC) separately. During the period between 2006 – 2017, a constant Hepatitis C dataset which comprise of 4962 patients with non-obtrusive strategies, for example, blood serum, for example, hemoglobin (HBGL), platelet check, white platelets (WBC) tally[21][22][23], and liver capacity tests [alanine aminotransferase (ALT), egg whites, alpha-fetoprotein (AFP), aspartate aminotransferase (AST), circuitous bilirubin, global standardized proportion (INR), prothrombin fixation (PC), all out bilirubin], clinical data [age, liquor utilization and tobacco utilization, weight record (BMI) and gender], lab tests [Anti-Nuclear Antibody (ANA), Baseline_PCR, Creatinine, Diabetes, Glucose, Thyroid-invigorating hormone test (TSH)], ultrasonography on liver and spleen alongside Transient Elastography to quantify liver firmness (LS), and a few factors, for example, AST-ALT proportion (AAR) and Fibrosis-4 list (FIB-4) were utilized to assess parallel order utilizing six calculations (Bayesian Network, Decision Tree, Naïve Bayes, Neural Networks, Random Forest and Support Vector Machine). The six-element choice techniques are p-esteem + Feature Subset Selection (CFS), Information Gain[24][25][26][27],

Principal Components Analysis, avaricious stepwise, Genetic calculation, and Particle Swarm Optimization. Among six calculations the Bayesian Network indicated the most elevated precision of 68.9%, followed by SVM (67.8%). Among six-component choice strategies and features, the p-esteem + CFS shows the 9 best-chose features, for example, Gender, Platelet, Albumin, Total Bilirubin, Baseline_PCR, Liver, Spleen, Stiffness, and prothrombin fixation individually

Agarwal G G et al.[28] [29]The past examinations indicated that among various HCV genotypes, the 1 and 3 genotypes show dominatingly in the whole of India, while 4 and 6 appeared in certain pieces of India individually. The Decision tree (DT) is utilized to group genotype a (1 to 6) and genotype 1b. Then again, the effective forecast of antiviral treatment in ceaseless Egyptian patients is dissected by DT and the significant indicator is discovered to be alpha-fetoprotein (AFP) level. Because of the previously mentioned genotype and AFP, the new examination is directed in the Microbiology branch of King George's Medical University, Lucknow. This medical clinic has an Integrated Counseling and Testing Center (ICTC) office and anti-Retroviral Therapy (ART) Center. This clinic gives HIV testing, linkages for clinical and psychosocial care, and guidance for people living with HIV contamination. During the period January 2007 – July 2008, an assent structure is gotten from a sum of 350 HIV-contaminated grown-ups going to the ICTC and ART Center to select them in the examination. A pre-planned proforma with 90 ascribes was utilized to meet the subjects about the clinical indications. Along these lines, the dataset of 350 perceptions with 90 ascribes was utilized to assess the presence of HCV as a double class name. The arbitrary backwoods from the R–bundle is utilized for the expectation of the model and the precision is discovered to be 98.3%

Hashem, Somaya et al. [17[31][32][33]Defines management of HCV contaminated patients can be observed by the evaluation of liver fibrosis arranging in Chronic Hepatitis C (CHC), to check the visualization of the malady, to build up ideal planning for treatment, and to anticipate the reaction to therapy separately. Even though liver biopsy has been utilized as a superior analysis technique, this strategy has potential dangers, for example, obtrusive nature, at risk of testing blunder, and the cost of observing. To conquer this, an option is non-intrusive strategies, for example, blood serum markers [alanine aminotransferase (ALT), egg whites, alpha-fetoprotein (AFP), aspartate aminotransferase (AST), Creatinine, glucose, hemoglobin (HB), backhanded

bilirubin, worldwide standardized proportion (INR), platelet check, postprandial glucose test (PC%), amount of HCV_RNA, serology discovering, absolute bilirubin, white platelets (WBC) count], alongside clinical data [such as age, sex, and weight file (BMI)), and contain histological discoveries, (for example, evaluation of fibrosis and the activity)]. Considering this, both serum biomarkers and clinical data have been contemplated to assess different AI procedures and create characterization models for the forecast of cutting edge fibrosis. Along these lines, four arrangement calculations, for example, choice tree, hereditary calculation, multi-direct relapse, and molecule swarm enhancement were utilized with Matlab and WEKA Software to assess the Cohort of 39,567 interminable HCV patients in Egypt. A sum of 21 ascribes are utilized for a double order. Among the four characterization calculations, the choice tree demonstrated the most elevated precision and AUROC of 84% and 0.76 separately and shows the four chose features age, AST, egg whites, and platelet tally which can be utilized for additional examinations

Hashem, Somaya, et al.[22] Use the Hepatitis C Virus dataset of the Egyptian patients for improving its accuracy using the Decision tree. The purpose of the study was to properly classify the data so the detection of the disease can be improved. For this purpose, they use different feature selection techniques for the selection of important features that are highly correlated to the output label. More correlated features will help in achieving more efficient incorrect detection of output labels. They implemented different types of decision trees such as classification and regression tree (CART), alternative decision tree, reduced error pruning tree (RET) for the classification purpose. They trained two models with the same decision three which is an alternate decision tree Model1 has six features that were highly correlated with the output label. Those features were age, alpha-fetoprotein (AFP), Platelet Count, Body mass index (BMI), Albumin, and aspartate aminotransferase (AST) which have the least *p*-value. And in the model2 they reduce two more features Body mass index (BMI) and Albumin from the six and then trained the model on the remaining four features. By using their proposed methodology they achieve the best model 86.2% NPV, 0.78 ROC, and 84.8% accuracy on the test set, better than the classical FIB-4 method.

Table 3.2: Literature Review of Supervised data classifiers

<table>
<tr><th>Year</th><th>Author</th><th>Dataset</th><th>Technique /Lang</th><th colspan="4">Accuracy</th></tr>
<tr><td rowspan="11">2020</td><td rowspan="11">Satish CR Nandipati et al.[17]</td><td rowspan="11">Hepatitis C virus (HCV) Egyptian</td><td rowspan="4">Python</td><td colspan="2">Binary class Label</td><td colspan="2">Multiclass Label</td></tr>
<tr><td>Number of Features</td><td>Accuracy</td><td>Number of Features</td><td>Accuracy</td></tr>
<tr><td>12</td><td>50.39%</td><td>12</td><td>24.15%</td></tr>
<tr><td>21</td><td>50.1%</td><td>21</td><td>24.44%</td></tr>
<tr><td>29</td><td>50.15%</td><td>29</td><td>24.8%</td></tr>
<tr><td rowspan="5">R</td><td colspan="2">Binary class Label</td><td colspan="2">Multiclass Label</td></tr>
<tr><td>Number of Features</td><td>Accuracy</td><td>Number of Features</td><td>Accuracy</td></tr>
<tr><td>12</td><td>50.59%</td><td>12</td><td>50.03%</td></tr>
<tr><td>21</td><td>49.57%</td><td>21</td><td>49.77%</td></tr>
<tr><td>29</td><td>50.9%</td><td>29</td><td>50.12%</td></tr>
<tr><td rowspan="2">2019</td><td rowspan="2">Abd El-Salam et al.[18]</td><td rowspan="2">Egyptian chronic hepatitis C patients</td><td>SVM</td><td colspan="4">67.8%</td></tr>
<tr><td>Bayesian Network</td><td colspan="4">68.9%</td></tr>
<tr><td>2019</td><td>Agarwal G G et al.[28]</td><td>Hepatitis C virus(HCV) Dataset from India</td><td>Decision Tree</td><td colspan="4">98.3%</td></tr>
</table>

2017	Hashem, Somaya et al. [30]	Egyptian chronic hepatitis C patients		Models		Accuracy	
				PSA		66.4%	
				GA		69.6%	
				MReg		69.1%	
				ADT		66.3%	
				ADT8		84.4%	
2016	Hashem, Somaya, et al.[22]	Egyptian chronic hepatitis C patients	Decision Tree Algorithms	Model1		Model2	
				No. of Feature	Accuracy	No. of Feature	Accuracy
				6	85.7%	4	85.7%

Abbreviations: PPV, positive predictive value; NPP, negative predictive value; ROC, receiver-operating characteristic curve; PSO, particle swarm optimization; GA, genetic algorithm; MReg, multi-linear regression; ADT, alternating decision tree. ADT* model with criteria point of zero.

3.3. CONVOLUTION NEURAL NETWORK (CNN)

Liu, Bing, et al.[34] Proposed a methodology which is the combination of the CNN model and Support Vector Machine (SVM) for the classification of the data which is in the form of images. CNN is very famous in the domain of hyperspectral image classification. But the CNN still facing challenges due to the less labeled training samples. The proposed features extracted technique depends just on a couple of boundaries, in particular, the learning rate, the edge boundary, and the number of ages. Furthermore, it is anything but difficult to set boundaries. The learning rate ought to be set to be a generally enormous worth (0.01 in this paper) toward the start of the preparation system. After a specific phase of preparing, a major rate may not, at this point be appropriate since it causes violates and give a higher misfortune. The learning rate is set to be 0.001 after 30 epochs. According to their proposed methodology, they used five different layers and use the output layer as the numeric feature extractor from the images dataset and then store these numeric values in the CSV file. And at the end, they use the Support vector machine (SVM) for the classification of the features. Thereafter, features extracted using S-CNN are utilized to prepare a direct SVM classifier. The test results with three notable hyperspectral informational indexes exhibit that the features extracted by S-CNN can incredibly improve the

presentation of hyperspectral picture grouping. Even with few preparing tests with the penance of expanded computational expense. According to the results, the combination of SVM and SCNN performed excellently better than that of the conventional methods.

Yang, Aimin et al. [35] proposed the classification technique based upon the CNN model for the medical-related data which is known as tumor images. First of all, they extract the features from the binary images of the tumor by rotation invariance. As the images shift rotation changes. The proposed methodology accurately defines the texture of the tumor image by the shallow layer. Furthermore to enhance the quality of the feature extraction through the CNN model which is already built and they break the limitation of the machine and human vision by using two models. The models which are used to improve the accuracy of the CNN model are Xception and Dense Net. By using the proposed methodology and the tumor images dataset they showed high accuracy in the classification of the tumor images. For the validation of their proposed methodology, they use three different datasets of tumor naming A, B, and C. And compare their accuracy result which is obtained by Xception and Dense Net with the Local binary patterns (LBP) algorithms like LBP_C, LBP_M, and LBP_S. The maximum accuracy obtained by the local binary patterns was 96.62% for A, 96.53% for B, and 87.12% for the C dataset. And the maximum accuracy was obtained by the Xception and dense Net 97.67% for A, 98.03% for B, and 91.6 for C dataset. By the results, you can see the difference between the accuracies of LBP and Xception and Dense Net with the combination of CNN models.

Chen, Yushi, et al.[36] Discussing the improvement of the efficiency of the hyperspectral images dataset. The aim of their proposed methodology was the increase the efficiency of the dataset which is a matter of concern these days. Due to less training data and high dimensionality, it's tough to do this. For the improvement of accuracy, they use the combination of the Convolution Neural Network (CNN) model and the Support Vector Machine (SVM). They used three-layer CNN with 4*4 or 5*5 convolution kernel and 2*2 pooling kernel. For the validation of the proposed methodology, they used three different datasets of the hyperspectral images which were Indian Pines, University of Pavia, and Kennedy Space Center. 1D CNN used for Spectral features, 2D CNN used for spatial features, and 3D CNN for both Spectral and Spatial features. Therefore they used 3D CNN for their approach. At the first, they use the CNN model for the feature extraction from the Images. And for the improvement of the efficiency of the model and avoid the over-fitting they use L2 regularization. Which used the Sum of the Square values and

smaller the value to avoid over-fitting. And after the feature extraction from the CNN model using L2 Regularization they used numeric classifiers like SVM for the classification of the extracted features. They achieved high accuracy with different datasets was 98.53% with Indian pines, 99.66% for the University of Pavia, and 97.07 for Kennedy Space Center datasets.

Table 3.3: A literature review of CNN Classifier with numeric classifier

Year	Author	Technique	Dataset	Accuracy	
2017	Liu, Bing, et al.[34]	S-CNN+SVM	Hyper-Spectral Images		
			Indian pines	99.14%	
			University of Pavia	99.08%	
			Kennedy Space Center	99.62%	
2019	Yang, Aimin et al. [35]	CNN+ SVM	Tumor Images Dataset	Xception	Dense Net
			A	97.67%	97.56%
			B	97.90%	98.03%
			C	91.6	90.14%
2016	Chen, Yushi, et al.[36]	CNN+ SVM with L2 Regularization	Hyper-Spectral Images from		
			Indian pines	98.53%	
			University of Pavia	99.66%	
			Kennedy Space Center	97.07%	

Lack of frameworks to input real data into deep learning model easily where dimensionality is high. A large number of datasets having poor accuracy in every field. Less efficiency of the supervised classifiers without the deep learning models for the complex dataset. Absence of automatic approach to classifying supervised data as images using deep learning models. These all are the main research gap which boosts us up to work in it.

CHAPTER 4. METHODOLOGY

The aim of the methodology part in the proposed study is to present the details and implementation of the Chernoff faces visualization techniques and the CNN model followed by the supervised classifiers like Support Vector Machine (SVM), Decision tree, Random Forest, Bagging, and KNN. This is a supervised machine learning algorithm for classification. Here below we have the flow diagram of our proposed methodology which presents the flow of the steps which we will follow for the validation of our proposed methodology. First of all, we need to collect the data to validate our proposed methodology. Data has more importance because it depends on what we are trying to do mean in which field and about what problem. It depends on the aim of the study about which about what we are going to solve. After the dataset collection, there is some pre-processing which is compulsory to do to bring the data into an understandable form. Because data gathered from different resources and can be in different formats and could have many anomalies and blunders. So the aim of the pre-processing is to purify the data from these anomalies and blunders. After the pre-processing, we need to collect those features which are important and playing important role in the output label. The importance of the feature depends on its contribution to the output label. So we would collect those features which have a high correlation to the output label. After this, we will move towards the main steps of our proposed methodology which is the visualization of data in form of Chernoff faces followed by the feature extraction from the images through the CNN model. After that, the last step of our flow is to classify the extracted features which are collected from the CNN model using different supervised algorithms, and finally, we will compare their accuracy with the result which was collected from the original dataset using the same supervised algorithms.

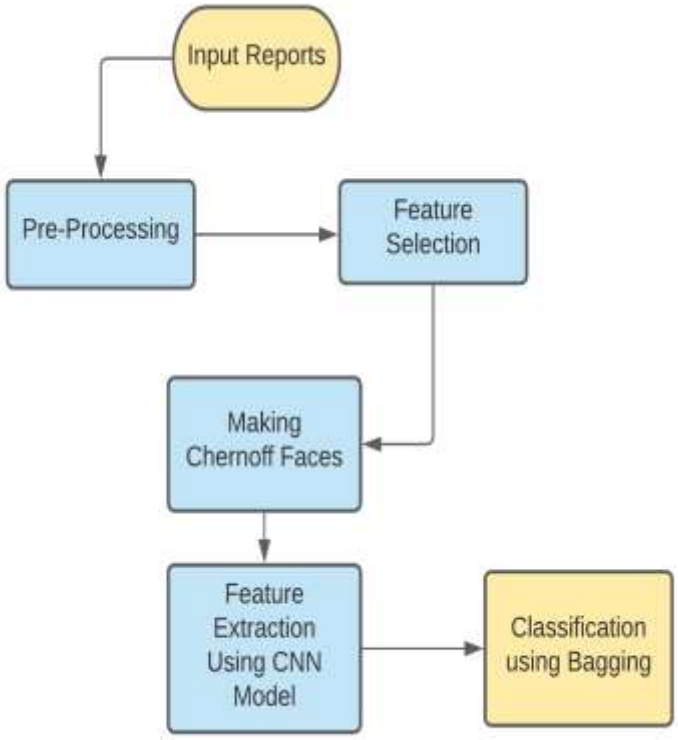

Figure 4.1: Flow of the Proposed Methodology

4.1. DATASET COLLECTION

The first and the main steps of the methodology is the selection of data on which we are going to implement our proposed methodology. The selection of the dataset depends on which field you are interested to do work or which specific area needs improvement. So our dataset collection aims to improve the accuracy of the Hepatitis C virus (HCV) which is very important for human life. A large number of peoples are affected by this disease in the whole world due to different causes. And lots of patients have died due to this deadly disease. Complete worldwide HCV commonness is assessed at 2.5% (177.5 million of HCV contaminated grown-ups), going from 2.9% in Africa and 1.3% in the Americas, with a worldwide pace of 67% (118.9 million of HCV RNA positive cases), differing from 64.4% in Asia to 74.8% in Australasia. HCV genotype 1 is the most common around the world (49.1%), trailed by genotype 3 (17.9%), 4 (16.8%) and 2 (11.0%). Genotypes 5 and 6 are liable for the remaining < 5%. While genotypes 1 and 3 are

normal around the world, the biggest extent of genotypes 4 and 5 is in lower-pay nations. Even though HCV genotypes 1 and 3 diseases are the most predominant universally (67.0% whenever thought about together), different genotypes are discovered all the more generally in lower-salary nations where represent a noteworthy extent of HCV cases[1]. In the images below we can see the effect of the Hepatitis C Virus is major in those countries which have low-income.

Figure 4.2: World Wide Effects of Hepatitis C Virus[37]

So the Aim out study to improve the accuracy of the supervised datasets and it will work for almost every supervised dataset with slight changes. So we select the Hepatitis C Virus (HCV) Egyptian Dataset to validate our study. It will help us to validate our study and with this, we tried our best to improve the efficiency of the dataset so it would be helping the doctor to analyze the patients correctly and will overcome the disease before the critical stages.

4.2. PRE-PROCESSING

After the selection of the dataset, the main step is the pre-processing of the dataset. In the field of machine learning, there are different steps which are fall in the pre-processing like data cleaning, data integration, data transformation, data reduction, data discretization, and data sampling. In the data cleaning, we fill the missing values and delete rows with missing data to resolve the inconsistency of data. In data integration, we put the data of different formats into a single

format. In the data transformation process, the data is normalized and generalized. The data reduction process includes the reduced presentation of data. Data Discretization involves the reduction of several values of a continuous attribute by dividing the range of attribute intervals. Some of the time, because of time, storage, or memory constraints, a dataset is too large or too complex to even consider being worked with. Inspecting strategies can be utilized to choose and work with only a subset of the dataset, given that it has roughly similar properties to the original one.

The dataset we used was very good. So with the slight changes, we were able to do further implementation on the dataset. In the original dataset they used male and female in the gender section and use present and absent in the Fever, vomiting, headache, Diarrhea, Fatigue and generalized bone ache, Jaundice, and Epigastric pain section. So we just replace the male with '1' and the female with '2'. And so on we represent the absent with '1' and present with the '2'. According to our requirement, this was necessary for us to bring the data into numeric form. Dataset was quite normalized and does not have missing or zero values. So the task of pre-processing was quite simple for us. In Table 4.1 we can see the data after preprocessing. We replace the string male and female with the numeric values with 1, 2 respectively in the gander attributes. And replace absent and present with 1, 2 accordingly.

Table 4.1: Dataset after Pre-Processing

Age	Gender	BMI	Fever	Vomiting	Headache	Diarrhea	Fatigue	Jaundice
56	1	35	2	1	1	1	2	2
46	1	29	1	2	2	1	2	2
57	1	33	2	2	2	2	1	1
49	2	33	1	2	1	2	1	2

4.3. FEATURE SELECTION

Feature selection is a very important step in the process of improving the accuracy of the data. If we are dealing with a high level of dimension then it is really difficult to improve accuracy. Feature selection plays an important role in improving accuracy. In feature selection, we select those features which are highly correlated to the output label. We have different feature selection

techniques for the selection of the best features Filter method, wrapper method, and Embedded method. We used the filter method and wrapper method for the feature selection process and with this, we used Correlation with heatmap. So we will discuss the filter method and wrapper method and in the end, we will discuss the Correlation Heatmap.

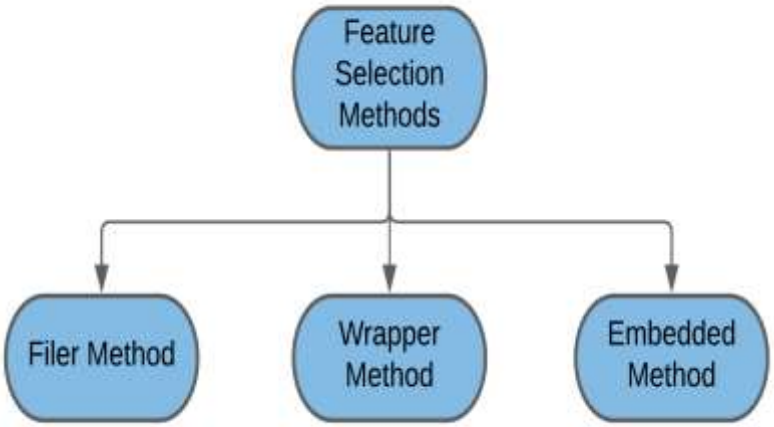

Figure 4.3: Feature Selection Techniques

4.3.1. FILTER METHOD

The filter method is general used as the preprocessing step of data. It is independent of any ML algorithm. The results are calculated using simple statistical scores which are used to determine the relation of the feature with the output label. The correlation coefficients for the different types of the dataset are the following show in Table 4.1 given below.

Table 4.2: Coefficient for Different Datasets

Feature/Response	Continuous	Categorical
Continuous	Pearson's Correlation	LAD
Categorical	Anova	Chi-Square

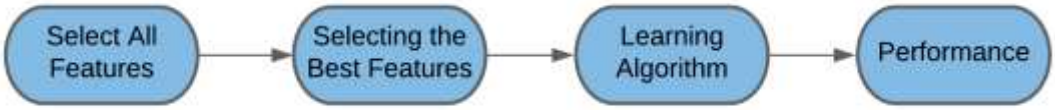

Figure 4.4: Filter Method for Feature Selection

Univariate selection is a statistical test that can be used to analyze which feature has the strongest relationship with the output label individually. For the Chernoff faces the total number of features required is eighteen but one feature we will keep the same which is the height of the upper face. So we will select the best seventeen features by using different feature selection techniques. In univariate we use the SelectKBest function with the score function of Chi-square and select 17 best features.

4.3.2. WRAPPER METHOD

The wrapper method is also a technique used for feature selection. This method needs a machine learning algorithm for the evaluation criteria Feature which is best for the ML algorithm and aims to improve the performance. To features evaluation, the accuracy used for classification tasks and goodness of cluster is evaluated using clustering.

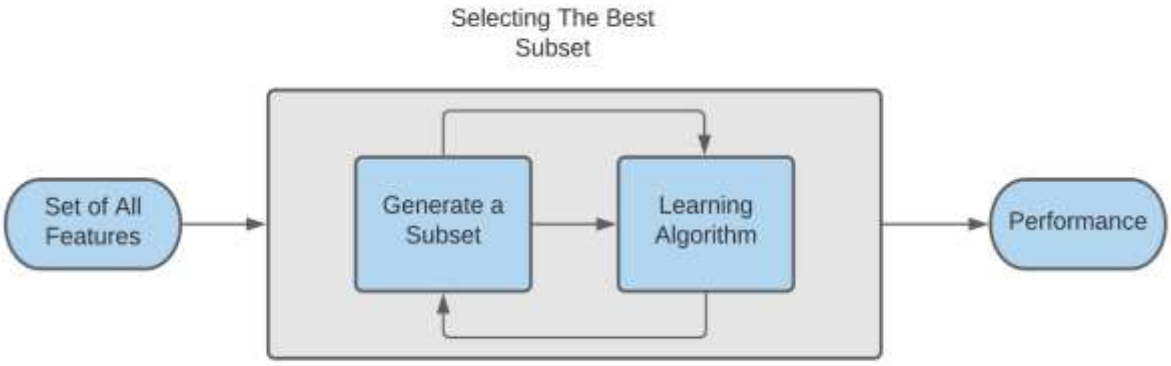

Figure 4.5: Wrapper Method for Feature Selection

We used feature importance which is also a type of wrapper method. This gives us the importance of the feature from the data. Features that have a higher score is a more important or relevant feature to the output. In the feature importance, we used ExtraTreeClassifier as the Machine learning algorithm and used its feature importance property for obtaining those features which have high importance towards the output.

4.3.3. CORRELATION WITH HEATMAP

After that, we used one more technique for the best confirmation of the best 17 features after that we can move on towards the implementation of our proposed methodology. So we used

correlation with heatmap for the priority matrix of n*n features. Values in the heatmap can be positive and negative as well. Features that have a high correlation will be represented by three colors red, yellow, and green. The priority value of each attributes depends upon its contribution to the output label. The priority value is between 0 to 1. An attribute having a priority value near to zero having less importance and the attribute having a priority value closer to 1 having much importance. So we will collect those features which will have the highest priority value among all the attributes.

With the help of these three visualization techniques, we select the best 17 features for the formation of Chernoff's faces.

4.4. MAKING CHERNOFF FACES

For the proper view of the data, we have many visualization techniques so we can understand data. And based on that understanding, we will be able to make a decision. For the visualization purpose, different visualization techniques are (scatter plot, histogram, Chernoff faces, Pixel-oriented visualization, Geometric Projection Visualization, Icon-Based Visualization, Hierarchical Visualization, Visualizing Complex Data, and Relations, etc.). All these visualization techniques have different purposes in different conditions. As the aim of our proposed methodology is to use the CNN model for the classification of the data efficiently without any blunders and difficulties, so we use the Chernoff visualization technique. A Chernoff's face is the multi-variant data visualization technique proposed by Herman Chernoff in 1973. That was developed for the visualization of the data in the form of a human face. Because each record represents an image this looks like the human face. And it's really easy to deal with the images for the CNN models without any difficulties. And so on CNN model will extract the important features from the images. Figure 4.6 is the little demo of the Chernoff faces visualization technique. Each face represents one record having 18 features. In the example, we can see twelve faces its means that is the representation of the twelve rows having eighteen features each. Through this visualization, we can see the difference between all faces. Some faces are small in size some have little mouth some have bigger eyes. Each feature of the face represents one attributes value from the dataset. Thus 17 attributes will represent in form of 17 features of the Chernoff. As we can see the faces below are much arranged. It looks like a proper face and we can see each feature of the face properly and on its slot. This is all about the

arrangement of the attributes. Assignment of the value of the attribute to the features of the faces does matter in the making of Chernoff faces in python. So while making off Chernoff faces for the Hepatitis C virus (HCV) data we arranged our attributes in a way that Chernoff's face looks good to see. The figure below this is just a sample of proper visualization of data and arrangement of attributes in the correct form to make a proper Chernoff face.

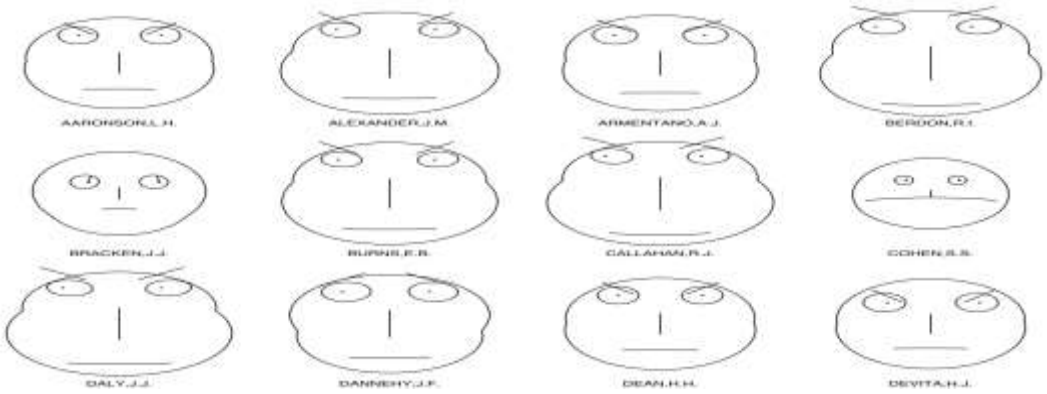

Figure 4.6: Chernoff Faces[2]

We are using matplotlib libraries for the making of Chernoff faces using python language. As the selection of the important features does matter then the arrangement of the attributes in the making of Chernoff faces in python also matters. If we are working will Rstudio then it is not a problem to deal with the arrangement of the attributes but in python, it does matter. If we are not properly arranging the attributes then the shapes of the faces will not be good.

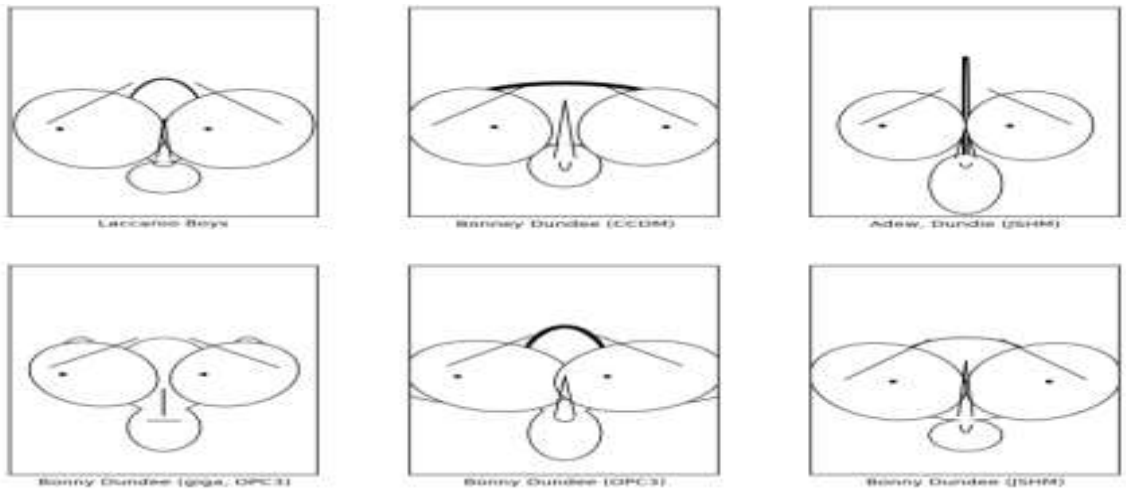

Figure 4.7: Bad Assignment of Features for Making Chernoff of Faces[38]

4.4.1. FEATURE DESCRIPTION

So we assign the attributes of the HCV dataset in a way to properly visualize each feature. So we have the 18 attributes helping in making Chernoff's faces. Explanations of the features are following

x1 = Upper face height

x2 = Lower face overlap

x3 = Vertical size of the half face

x4 = Upper face width

x5 = Lower face width

x6 = Nose length

x7 = Mouth vertical position

x8 = Mouth curvature

x9 = Mouth width

x10 = Eyes vertical position

x11 = Eyes separation

x12 = Eyes slant

x13 = Eyes eccentricity

x14 = Eyes size

x15 = Pupils position

x16 = Eyebrows vertical position

x17 = Eyebrows slant

x18 = Eyebrows size

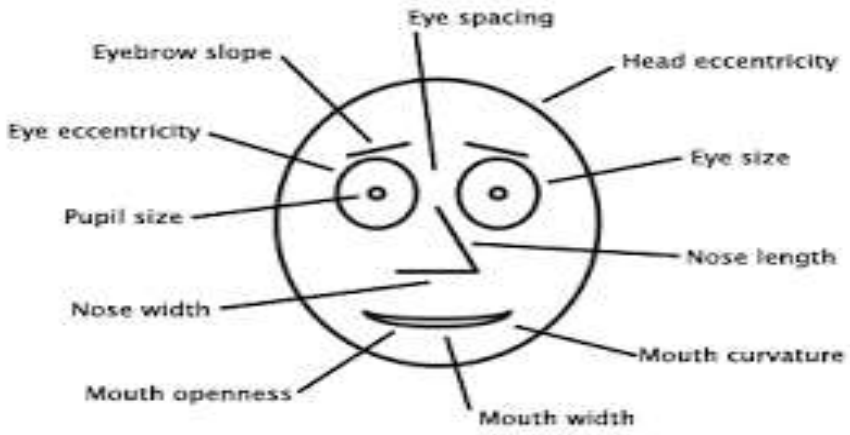

Figure 4.8: Good Feature Assignment for Chernoff Faces[39]

Here is the sample of Chernoff's face for assigning the attributes. And explain the features of Chernoff's face like how we will calculate and the values to the Chernoff's face.

And here is the function below we used for the making of Chernoff's faces and the arrangements of the attributes which we used for the making of Chernoff's faces.

def cface (ax, x1, x2, x3, x4, x5, x6, x7, x8, x9, x10, x11, x12, x13, x14, x15, x16, x17, x18):

And here is the arrangement of the 17 attributes into the function for the proper visualization

cface(ax, 0.9, cv1.WBC, cv1.BMI, cv1.RBC, cv1.Plat,cv1.AST_1, cv1.Age, cv1.ALT_1, cv1.HGB, cv1.ALT_4, cv1.ALT_48, cv1.ALT_12, cv1.ALT_36, cv1.ALT_after_24w, cv1.RNA_Base, cv1.RNA_4, cv1.RNA_EF, cv1.Baseline_histological_Grading)

And further, we have the normalization formula for the attributes

Def NormalizeData (data):

return (data - np.min (data)) / (np.max (data) - np.min (data))

After these, we plot the Chernoff faces for the attributes of each file and loop over all four files. We got a single image for each CSV file which was a problem. Because we need the number of images for each file is equal to the number of records. But we need 336 images for the extraction of the features by using the CNN model for the 336 records but we got just one image of all

Chernoff faces. So slice the single image into multiple to get the number of images is equal to the number of records in each file.

4.5. FEATURE EXTRACTION USING CNN MODEL

In this process, we aimed to extract the important features from the Images that were produced by the original Hepatitis C virus (HCV) dataset with the help of Chernoff faces techniques. For this purpose, we used a Sequential Convolution neural network with multiples Convolution layers discussed below. We will discuss the libraries and properties that we used for the implementation of our proposed technique. First of all, we will discuss the libraries which we used for importing some essential properties for python. Figure 4.10 is the details of the convolution layers which we used for the process of the feature extraction. For this, we used 14 different layers to build a sequential Convolution neural network for the feature extraction. We used five Conv2D, five Maxpool2D, Three Dense, and 1 Flatten layers. These layers are arranging to build the sequential CNN model. We used a Conv2D layer with a kernel size of 3x3 and have a filter size of 8 and the size of the original image was 183x183 so after valid padding of the image becomes 181x181 and we used a filter size of 8 for the first conv2D layer. A second layer is Maxpool2D which reduced the size from 181x181 to 90x90 and after valid padding, it becomes 88x88. So the input for the third layer which is con2D again is 88x88 with the kernel size of 8 which means it divides the image into 8 sub-parts. The fourth layer is again Max2D which reduced the size to half of the original and provide the size of 44x 44 after padding it remains 42x42 which is the input for the fifth layer. And it reduced the image further into 8 sub-parts. Furthermore, the sixth layer is again Max2D which reduced the size to 21x21 with the kernel size of 8, and after padding, it remains 19x19 and that is the input for the seventh layer which is the Con2d layer. The eighth layer reduced the size from 19x19 to 9x9 and after padding, it remains the 7x7 with the kernel size of 8 which becomes the input for the ninth layer of Conv2D. Furthermore tenth the Maxpool2D layer reduced the size from 7x7 to 3x3 with the kernel size of eight. After that, we used 3 consecutive dense layers having a size of 4,3x3, 4,3x3, and 3,3x3 accordingly. After the 3 Dense layers, we have a flattened Layer that extracts the most important 27 features for each image.

This process is loop over all four folders having Chernoff faces with different output labels. The extracted features are stored in different CSV files. After that, we label all the CSV files with the

original output label and combined them into a single file manually. After that, we will be able to classify the dataset of those features which we extracted from Chernoff's faces.

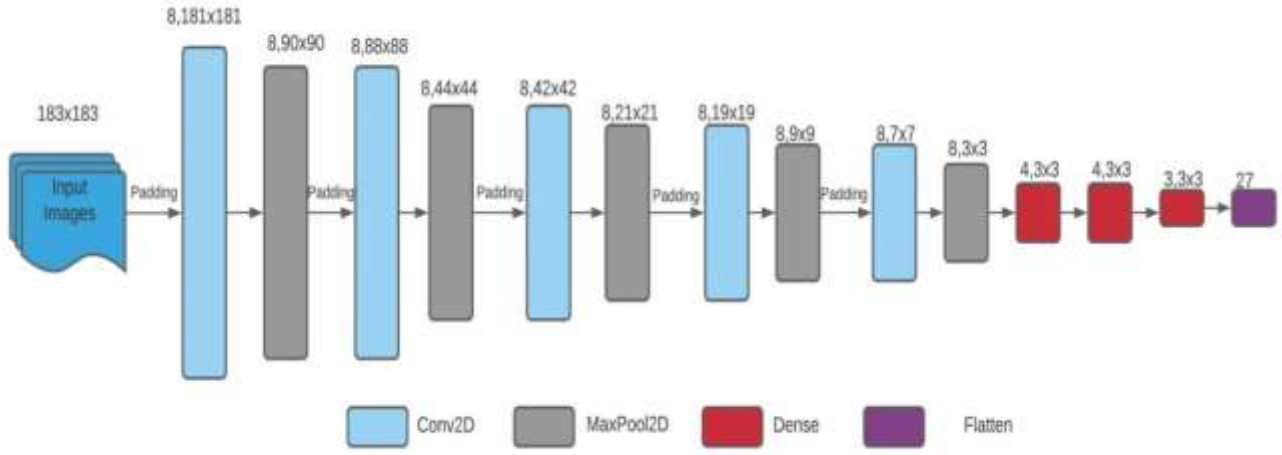

Figure 4.9: CNN Model for The Features Extraction

4.5.1. LIBRARIES

For this purpose, we need some libraries to import into pycharm for the implementation of the CNN model in Python. Like

From os import listdr

From keras.preprocessing.image import load_img

From keras.preprocessing.image import img_to_array

From keras.models import sequentially

From keras.layers import Dense, Activation, Flatten, Conv2D, MaxPool2D

We used listdir () that is used to return the list of the name of the entity from the directory by the given path. It does not include the special entries '.' and '.' even if they are present in the directory. A path can be either type bytes or type string.

After that, we used the load_img property of keras.processing.image to load that image. Load_img contains different properties like file path, which is the actual path of the image. Then grayscale which is the deprecated use color_mode= “grayscale”. Then it has a color-mode which is used to get the desired grayscale image format like” rgb”, “rgba”, Default: “rgb” color. Target_size which is used to specify the size of the image in our case target size was (183,183).

And at the end interpolation is used to resample the image if the target size is different from that of the load image.

Furthermore, we have the img_to_array property which is used to convert the image into the 3D NumPy array. It has different parameters like img, data_format, and data_type. Img contain the actual load image, data_format of an image can be either ‘channel_first’ or ‘channel_last’. By default, it is ‘None’. In the end, we have keras. model and import its sequential property which is our CNN model used for feature extraction and its parameters like Dense, Activation, Flatten, Conv2D, and MaxPool2D.

4.5.2. CONVOLUTION NEURAL NETWORK

CNN has altogether dominated the machine vision space in recent years. These neural nets are so amazing and persuasive that they have made profound learning probably the most sultry subject in Artificial Intelligence (AI) today. Yann Lecon of New York University, who additionally fills in as the director of Facebook's AI gathering, pioneered CNNs. They are a sort of profound neural networks that were planned from the naturally determined models. Researchers found out that mammals or humans visually perceive the surrounding world using a layered architecture of neurons in the brain. This, in turn, motivated engineers to design similar pattern recognition systems and that is how these nets were evolved.

4.5.3. RESEMBLANCE WITH MLP

The idea of CNN is the same as of multilayer perceptron (MLP). Cutting hardly any associations and sharing loads of MLP brings about one CNN layer as appeared in Figure 4.10. The figure portrays that such a counterfeit neural organization (ANN) setup is indistinguishable from 2D convolution activity and loads are simply channels (likewise called covers or parts). A major advantage of sharing inclinations and loads is that it fundamentally diminishes the no. of boundaries associated with a convolutional network.

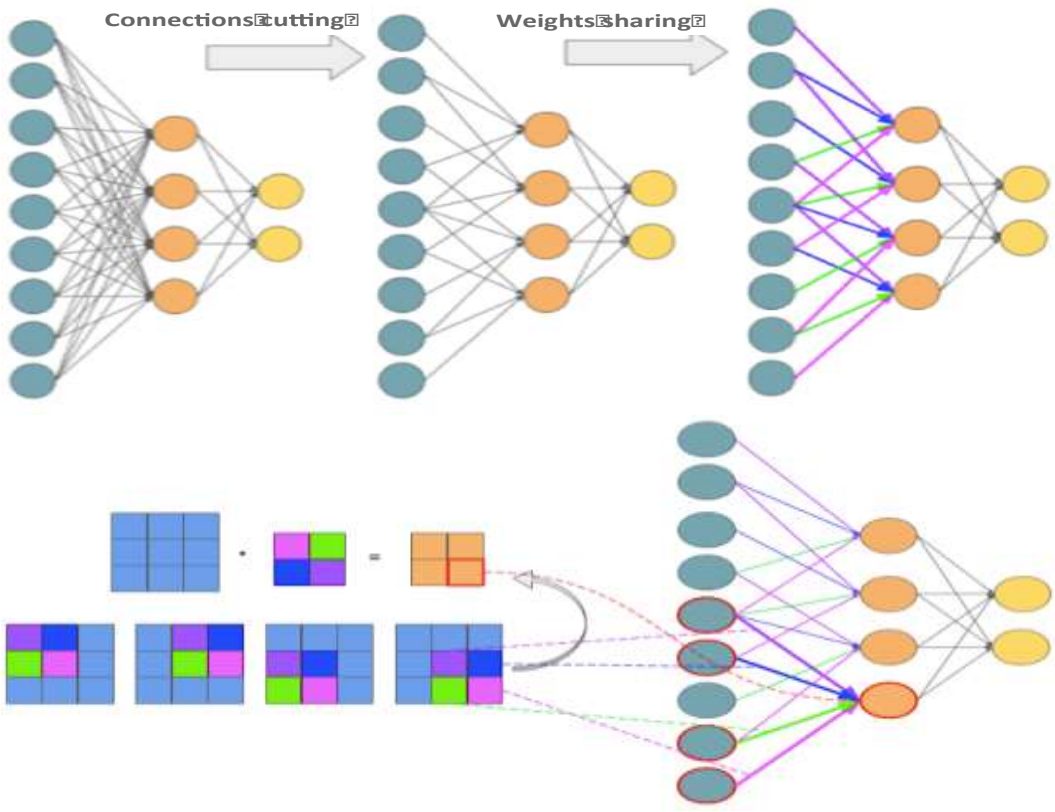

Figure 4.10: Structure of the CNN Model

4.5.3.1. ARCHITECTURE

A CNN incorporates data, various concealed layers, and a yield layer. The concealed layers normally involve convolutional layers, enactment layers, pooling layers, and fully connected (FC) layers. The initial three layers are applied reliably in a consistent movement and they help in features learning while the last FC layer is used for the grouping purpose.

Following layers are the main building blocks of a typical CNN:

4.5.3.2. CONVOLUTION LAYER

The convolutional layer comprises a bunch of free portions or channels and each channel is independently convolved with the figure. The convolution is finished by taking a channel, sliding it over the total picture, and en route taking the dab item between lumps of the picture and the channel. All the channels are introduced arbitrarily and they are the boundaries that will be found out by the organization thusly. The underlying layers search for essential examples, for example, lines or corners. As we go further to other convolutional layers, the channels are doing spot items

to the contribution of the last convolutional layers. Along these lines, they are taking the removed edges and make bigger parts out of them.

4.5.3.3. ReLU (RECTIFIED LINEAR UNITS) LAYERS:

After each convolutional layer, it is conventional to apply a nonlinear layer (or initiation layer) quickly thereafter. This layer targets acquainting nonlinearity with a framework that has quite recently been registering direct tasks during the convolutional layers (just component astute duplications and summations). ReLU layers accelerate the organization's preparation (because of computational productivity) without rolling out a huge improvement in the exactness. It additionally assists with mitigating the disappearing angle issue, which is where the lower layers of the organization train gradually because the slope diminishes dramatically through the layers. ReLU applies the actuation work f(x) = max(0, x) to the entirety of the qualities in the info volume. As such, the layer just changes all the negative qualities of the zero. It builds the nonlinear properties of the model and the general organization without upsetting the open fields of the convolutional layer.

4.5.3.4. POOLING LAYER

Pooling layers likewise alluded to as down-inspecting layers, focuses on continuously diminishing the spatial size of the portrayal in this manner decreasing the no. of calculations and boundaries in the organization. A few non-straight capacities can actualize pooling, for example, normal pooling, L2-standard pooling yet max pooling is the most well-known. Max pooling separates the picture into a bunch of non-covering lumps and, for each such piece, chooses the greatest worth. This way it decreases the spatial measurement (the length and the width change yet not the profundity) of the information volume. The thought behind pooling is that once we recognize the presence of a particular element in the first contribution (there will be high initiation esteem), its precise area isn't as fundamental as its general area to different features. So the pooling layer fills two fundamental needs. The first is that it decreases the calculation cost by lessening the number of boundaries or loads by 75%. The second is that it will power over-fitting. This term alludes to when a model is so tuned or fitted to the preparation models that it can't sum up well for the approval and test sets. The figure shows a case of how max and normal pooling is done.

4.5.3.5. FULLY CONNECTED LAYER

The convolutional, ReLU, what's more, pooling layers are applied again and again individually, lastly, the significant level thinking in the CNN is done through FC layers. The yield from the convolutional layers speaks to elevated level features in the picture and adding an FC layer permits you a non-direct mix of those features. All the features from convolutional layers might be acceptable (expecting we don't have "dead" features), however, mixes of those features may be far better. Neurons in an FC layer have full associations with all enactments in the former layer, as found in standard ANNs, and work correspondingly.

By utilizing the various libraries and numerous convolution layers we removed the best features from Chernoff's appearances. Also, annex them into various CSV records as per the yield mark. Moreover, we move towards the cycle of grouping utilizing the separated component utilizing the Convolution Neural Network.

4.6. CLASSIFICATION

In this process, we utilized different classifiers for the classification of the features that are separated collected by the sequential CNN model. This cycle is likewise applied to the first dataset which we have before making the Chernoff faces. We utilized the various types of the first dataset and afterward figure its outcomes utilizing various classifiers like KNN, Naïve Bayes, Decision Tree, Support Vector Machine (SVM), Random Forest, and so on. Most importantly, we utilized these classifiers and compute the outcomes utilizing the total dataset. What's more, from that point forward, we select the best 17 features which we utilized for making the Chernoff faces. And afterward, we applied every one of these classifiers to check the precision of the dataset having 17 features. Furthermore, the third time we utilized these classifiers for the arrangement of the dataset which incorporates the features extricated from the Sequential CNN model. Besides, we will think about the consequence of the classifier utilizing three types of datasets. Here is the consequence of every one of the three types of dataset.

Above all else, we gather the dataset into three structures or dependent on the number of qualities. Three datasets are the following

1. Original Dataset
2. Selected Feature Dataset
3. Extracted Feature Dataset

We used different classifiers for the classification of these datasets. And then collect their results and show some images of the classifiers having good results. The original dataset has 28 attributes including the output label. The selected features dataset have 17 features that we select for making Chernoff's faces. And the extracted features dataset has 27 attributes extracted using the CNN model from Chernoff's faces. We used these three forms of datasets and show the results of four classifiers which are Random Forest, Decision Tree, Support Vector Machine (SVM), and Bagging. We used multiple algorithms but will just discuss the four which gave us the best results.

4.6.1. DECISION TREE

A Decision Tree (DT) is a non-parametric supervised learning technique used for gathering and relapse. Choice trees gain from data to infer a sine bend with a ton of if-else decision guidelines. More significant the tree, the more many-sided the decision rules and the fitter the model.

The choice tree builds portrayal models as a tree structure. It isolates an educational assortment into more unobtrusive and more humble subsets while at the same time a connected choice tree is consistently developed. The inevitable result is a tree with decision center points and leaf center points. A decision center has in any event two branches. Leaf center point addresses a portrayal or decision. The most elevated choice hub in a tree that thinks about the best pointer is called the root hub. Choice trees can manage both absolute and numerical data. Here is the model beneath a decision tree.

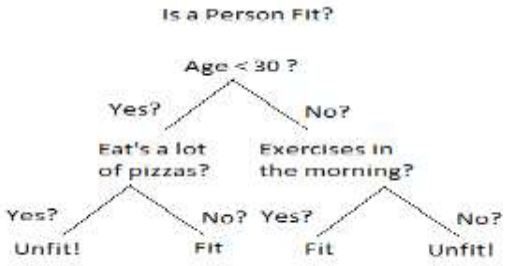

Figure 4.11: Decision Tree Classifier[40]

There are two approaches used in the decision tree are splitting and pruning and some factors involved in the decision tree are entropy and information gain

4.6.2. RANDOM FOREST

Random forest is a supervised learning calculation. It might be used both for characterization and regression. It is similarly the most versatile and easy to use count. A backwoods is contained trees. It is said that the more trees it has, the fierier a timberland is. Self-assertive forest areas settle on choice trees on heedlessly picked data tests, gets conjecture from each tree, and picks the best plan by strategies for casting a ballot. It also gives an exceptionally nice marker of the component importance.

The arbitrary backwoods has a collection of uses, for instance, recommendation engines, picture gathering, and highlight determination. It might be used to arrange resolute development applicants, perceive bogus development, and predict contaminations. It lies at the base of the Boruta calculation, which picks huge features in a dataset. Here is the example below of the random forest classifier.

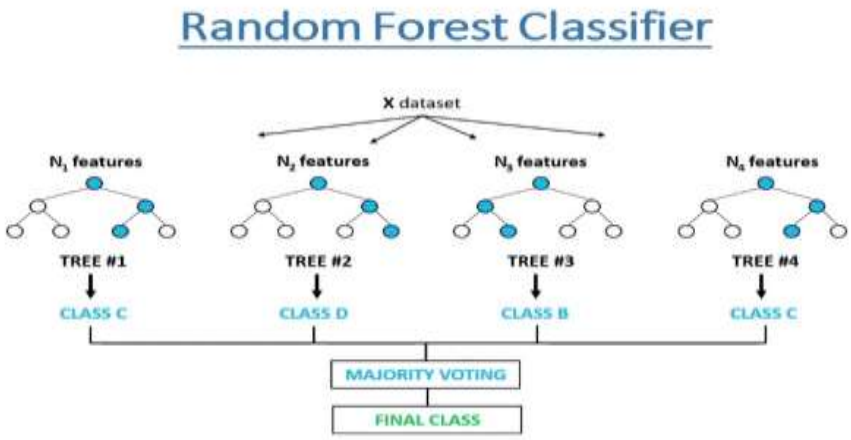

Figure 4.12: Random Forest Classifier[41]

4.6.3. SVM

SVM is an administered ML algorithm that is used for regression and, all the more frequently, for classification issues. In this calculation, every information esteem is plotted as a point in n-dimensional space (where n is the number of features) with the estimation of each element of a specific arrange[42][43]. At that point, the classification finds the hyper-plane that isolates the two classes better and amplifies the edge (for example the separation among it and the closest information purpose of each class). Figure 4.13 shows a few instances of how the calculation picks the privilege hyper-plane. In Figure 4.13-(A), hyper-plane B is distinguished as the privilege hyper-plane as it isolates the two classes better. In Figure 4.13-(B), all the hyper-planes

(A, B, and C) are isolating the classes well. For this situation, the hyper-plane having the most extreme edge from the closest information point will be chosen, consequently, hyper-plane C will be picked. In Figure 4.13-(C), B is a superior classifier yet SVM chooses the hyper-plane that orders the classes precisely before augmenting edge. Thus, the privilege hyper-plane is A, as it has no grouping mistake. SVM can likewise disregard the anomalies and amplify the edge as appeared in Figure 4.13-(D). Till now, we have only visualized the linear kernel but SVM can also solve a linearly inseparable problem using complex kernels e.g. radial basis function (RBF) kernel, a polynomial of higher degree, etc. Figure 4.13-(E) shows a circular kernel for the data that is not linearly separable. For Tuberculosis detection, linear SVM was used to project the ensemble features to high dimensional space and then finding the optimal hyper-plane.

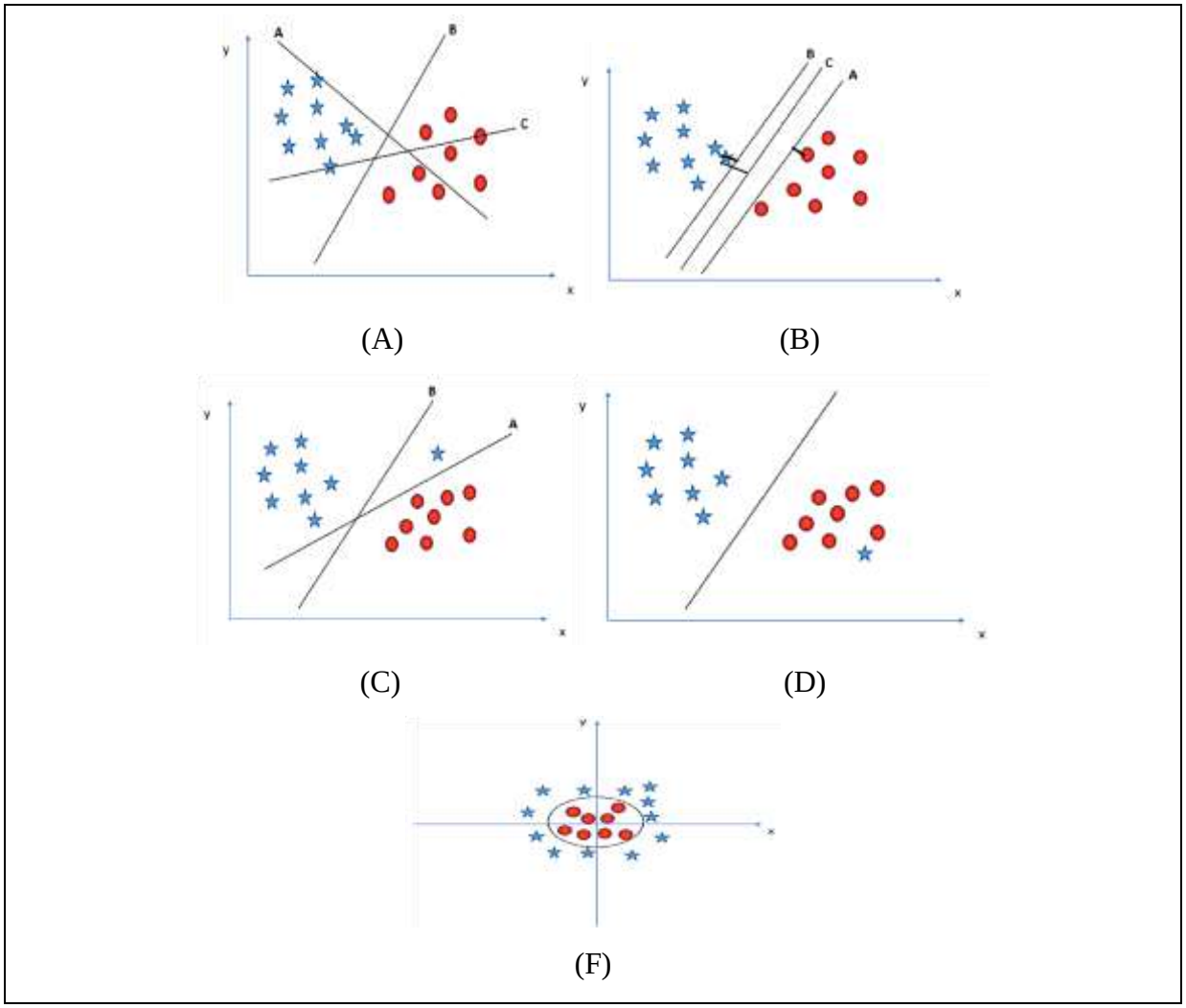

Figure 4.13: SVM Different Hyper Plans

4.6.4. BAGGING

Bootstrap aggregating also called packing (from bootstrap collecting), is an AI meta-calculation planned to improve the sufficiency and exactness of AI calculations used in the real course of action and relapse. It furthermore diminishes change and helps with avoiding overfitting. Even though it is regularly applied to decision tree methodologies, it will in general be used with a strategy. Stowing is a novel occurrence of the model averaging approach.

4.7. SUMMARY

Here is the detailed introduction of our proposed methodology above. The main focus of our proposed methodology is to improve the efficiency of supervised data via the CNN model. For this purpose, we used the Chernoff faces technique which made us able to do this. Using Chernoff faces visualization technique we visualize our supervised data in form of faces which is a set of images. And later on, we used these images as the input of the CNN model. And then extract 27 features from the CNN model by removing flatten later. And then we used different supervised classifiers to classify these extracted features. By following the whole process we achieve a good result. This helps us in improving the accuracy of the supervised data.

CHAPTER 5. RESULTS AND DISCUSSION

Here in this portion, we will discuss the results that we achieved by using our proposed methodology. The discussion will be about the techniques which we used and also about the dataset. We will also discuss the result we achieved before and after the proposed methodology. We will discuss the results a little bit about the proposed methodology and dataset. The sequence of the discussion is as follow dataset, Preprocessing and features selection, Chernoff faces, feature extraction using CNN, Classification. This is the flow of our discussion in the portion of results and discussion. The first part of the discussion is a dataset that is very important for the validation of the proposed methodology.

5.1. DATASET

Dataset is important for the validation of the proposed methodology. Our proposed methodology is used to increase the efficiency of the dataset. So the improvement in the dataset can help its stakeholders. And it has also important in its field. In that case, we deal with the medical-related dataset which is very crucial. The dataset that we used is the Hepatitis C Virus (HCV) dataset. This is medical related dataset and also very important. HCV is a very dangerous disease that has been spread worldwide. There are several cause factors for the spread of HCV. The improvement in the efficiency of that dataset can help in figuring out the level or stage of the disease of the patient and timely medication of that patient can help in saving a life. A lot of research has been carried out on this dataset in previous years for the improvement of the efficiency of that dataset. The maximum accuracy achieved for this dataset was 85.7%. That is a good percentage as well to achieve for a very complex dataset. The dataset of HCV which, we used is the Egyptian HCV dataset. This dataset has Twenty-nine attributes including output labels.

5.2. DATASET DIVISION

We divide the dataset into different CSV according to the number of output labels. We have four labels to identify the images after making Chernoff's faces. So we divide it into four files for the making of Chernoff faces for each file. After this, we used different matplotlib properties like matplotlib.Patches.Ellipse to make each feature of the face. After that, we supplied the value of each attribute into the function which was built for the making of Chernoff's faces and we also

normalized the value of the attribute so the features do not look awkward. The division of the dataset aims to obtain the separate data of each label. And build Chernoff faces of that specific dataset and then extract the features from those images using the CNN model. And in this way, we can keep track that which image and extracted features belong to which output label. Here is the chunk of code in the form of an image for dividing the dataset according to the output label. In which we divide the dataset into four different CSV's on the base of the output label and further we store the records of each output label into a different CSV file.

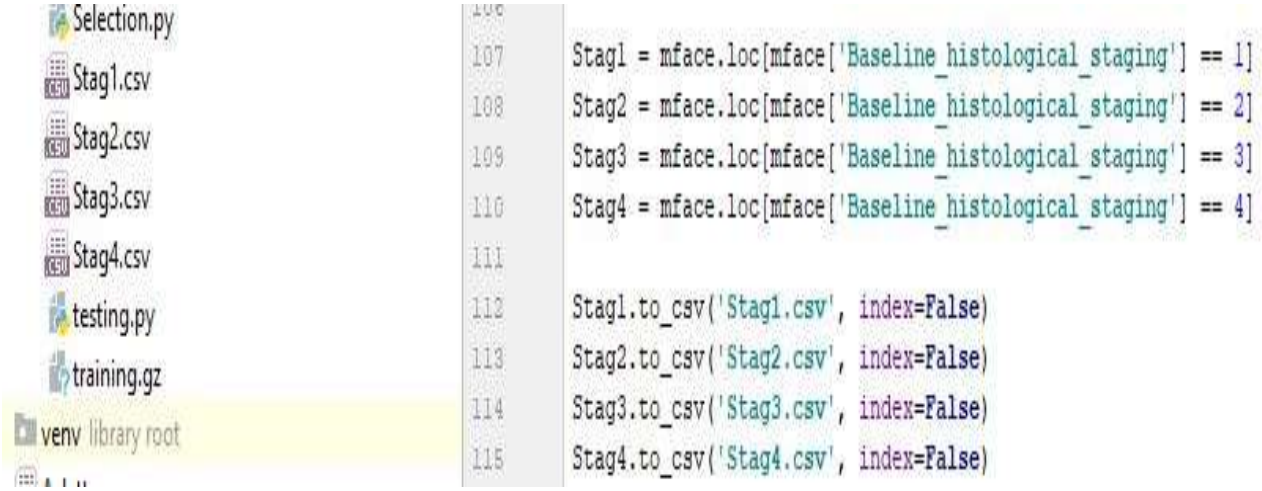

Figure 5.1: Dataset Division Code

5.3. PRE-PROCESSING AND FEATURE SELECTION

In this process, different techniques are used for the preprocessing and the feature selection of the features. We have 29 attributes including output labels and a little bit of preprocessing has been done of this dataset to bring it in data form which is useful for our proposed methodology. Dataset has some string values for some attributes but we need numeric values for the processing of making Chernoff's faces. Few changes have been done to the dataset for use. In the original dataset they used male and female in the gender section and use present and absent in the Fever, vomiting, headache, Diarrhea, Fatigue and generalized bone ache, Jaundice, and Epigastric pain section. So we just replace the male with '1' and the female with '2'. And so on we represent the

absent with ‘1’ and present with the ‘2’. According to our requirement, this was necessary for us to bring the data into numeric form. Dataset was quite normalized and does not have missing or zero values. So the task of pre-processing was quite simple for us. In Table 5.1 we can see the data after preprocessing. We replace the string male and female with the numeric values with 1, 2 respectively in the gander attributes. And replace absent and present with 1, 2 accordingly.

Table 5.1: Dataset after Pre-Processing

Age	Gender	BMI	Fever	Vomiting	Headache	Diarrhea	Fatigue	Jaundice
56	1	35	2	1	1	1	2	2
46	1	29	1	2	2	1	2	2
57	1	33	2	2	2	2	1	1
49	2	33	1	2	1	2	1	2

Here is the chunk of the dataset after the preprocessing in which we replace the string values with the numeric values. After the preprocessing, we used different feature selection techniques that are necessary for feature reduction which is important for the implementation of our proposed methodology. In our proposed methodology, we need seventeen features in the making of Chernoff's faces. In the original dataset, we have 28 features excluding the output label. And we need 17 features for Chernoff's faces. For feature selection, we used three different techniques which are the following.

- Uni-Variant Selection
- Feature importance
- Co-relation with heatmap

These three approaches we used for the selection of seventeen features.

5.3.1. UNIVARIATE SELECTION

Univariate selection is a statistical test that can be used to analyze which feature has the strongest relationship with the output label individually.

Table 5.2: Best 17 Features Using Univariate Selection Technique

Feature Number	Feature Name	Accuracy Score
26	RNA_EF	1.8
22	RNA_Base	6.33
24	RNA_12	5.5
23	RNA_4	5.03
25	RNA_EOT	1.78
11	RBC	4.20
13	Plat	1.81
10	WBC	1.22
15	ALT_1	2.07
14	AST_1	1.99
20	ALT-48	1.34
16	ALT_4	1.14
2	BMI	7.00
19	ALT_36	6.39
0	Age	6.18
18	ALT_24	5.40
27	Baseline_histological_Grading	5.35

Here is the result of the univariate selection which gives us the best 17 features in Table 5.2

5.3.2. FEATURE IMPORTANCE

This is a type of wrapper method in which we used a machine learning algorithm for the evaluation of the priority value of each attribute. Feature importance is the property of the model. This gives us the importance of the feature from the data. Features that have a higher score that is a more important or relevant feature to the output. In the feature importance, we used ExtraTreeClassifier as the model and used its feature importance property for obtaining those features which have high importance towards the output.

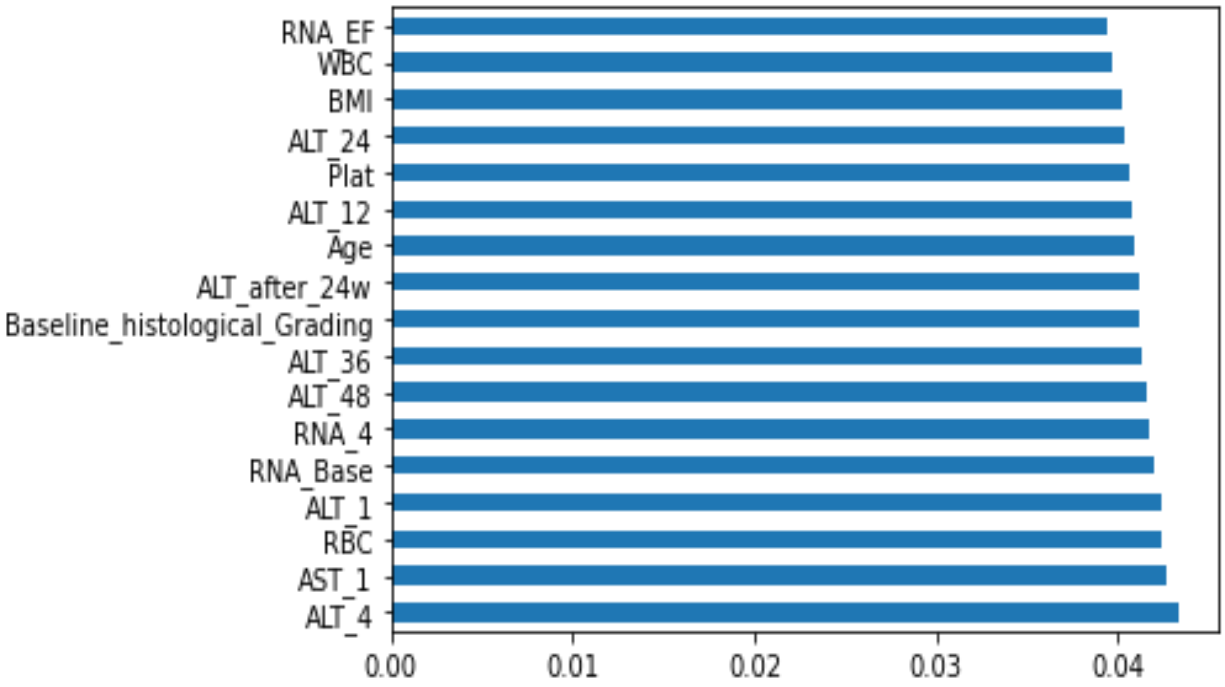

Figure 5.2: Best 17 Features Extracted Using Feature Importance

5.3.3. CORRELATION WITH HEATMAP

After that, we used one more technique for the best confirmation of the best 17 features after that we can move on towards the implementation of our proposed methodology. So we used correlation with heatmap for the priority matrix of n*n features. Values in the heatmap can be positive and negative as well. That feature will have a high correlation that will be represented by three colors red, yellow, and green. The priority value of each attributes depends upon its contribution to the output label. The priority value is between 0 to 1. An attribute having a priority value near to zero having less importance and the attribute having a priority value closer to 1 having much importance. So we will collect those features which will have the highest priority value among all the attributes.

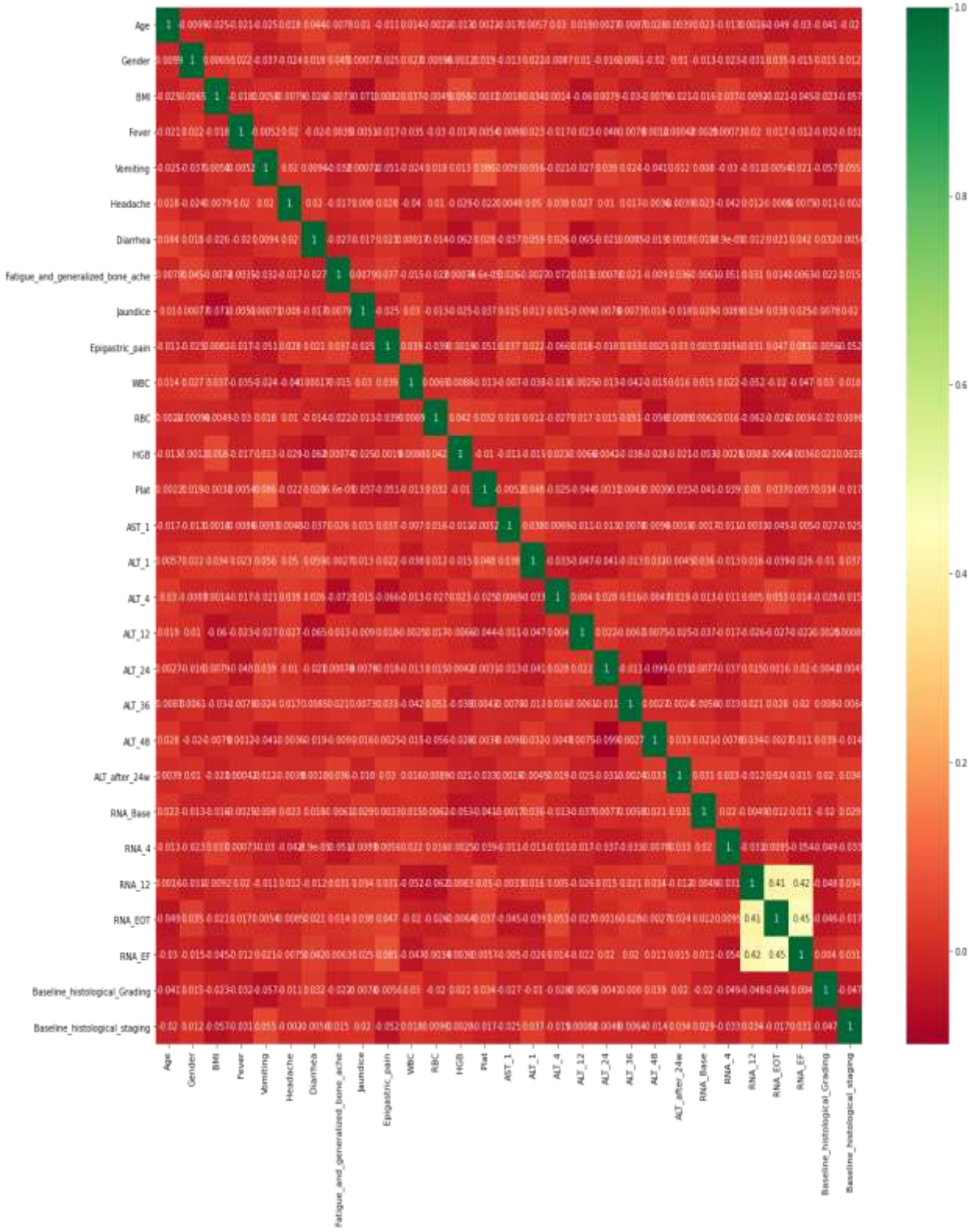

Figure 5.3: Priority Matrix of n*n Feature for Having 17 High Priority Features

After the implementation of these three techniques, we select seventeen best features that are high in priority according to the output label.

The features we selected after the feature selection process are the following WBC(White blood cell), BMI(Body Mass Index), RBC(red blood cells), Plat(Platelets),AST_1(aspartate transaminase ratio), age, ALT_1(alanine transaminase ratio 1 week) , HGB(Hemoglobin), ALT_4(alanine transaminase ratio 4 week), ALT_48(alanine transaminase ratio 48 WEEK), ALT_after_24w(after 24 w alanine transaminase ratio 24 weeks), RNA_Base, RNA_4, RNA_EF(RNA Elongation Factor), Baseline_histological_Grading. After the feature selection process, we can continue with the implementation of our proposed techniques. So we collect the best 17 features for the visualization of these features for the formation of Chernoff's faces.

5.4. CHERNOFF FACES

Here is the discussion about Chernoff's faces. After the process of feature selection, we select the seventeen features, and then we used those seventeen features for the formation of Chernoff's faces. At the first after the feature selection, we divide the dataset into four parts according to the number of output labels. So we can know to which output label Chernoff's face belongs. Chernoff's faces have a crucial position in the proposed methodology. We assigned seventeen attributes to different seventeen facial features. The assignment of the attributes to the facial features does matter in the proper formation of Chernoff's faces. We assign different attributes to different features as follows.

5.4.1. ATTRIBUTES ASSIGNMENT

So our assignments of the attributes to the Chernoff faces features are the following. We just fix the value of x1 manually

0.9 = x1 =>Upper face height
WBC = x2 => Lower face overlap
BMI = x3 => Vertical size of the half face
RBC = x4 => Upper face width
Plat = x5 => Lower face width

AST-1 = x6 => Nose length

age = x7 => Mouth vertical position

ALT_1 = x8 => Mouth curvature

HGB = x9 => Mouth width

ALT_4 = x10 => Eyes vertical position

ALT_48 = x11 => Eyes separation

ALT_12 = x12 => Eyes Slant

ALT_36 = x13 => Eyes eccentricity

ALT_after_24w = x14 => Eyes size

RNA_Base = x15 => Pupils position

RNA_4 = x16 => Eyebrows vertical position

RNA_EF =x17 => Eyebrows slant

Baseline_histological_Grading=x18 => Eyebrows size

5.4.2. MAKING CHERNOFF FACES

After the selection and the alignment of these attributes we move towards the making of Chernoff faces for the visualization of numeric data. We use different libraries in the formation of the Chernoff faces like matplotlib and its properties like matplotlib. use (‘Aug’) and to plot the Chernoff faces we use matplotlib. pyplot.

This was the assignment of the attributes to the facial features. After this, we used different matplotlib libraries and several properties for the mathematical form of Chernoff's faces. After the formation of Chernoff's faces, we got a single image for all the records as shown in Figure 5.4

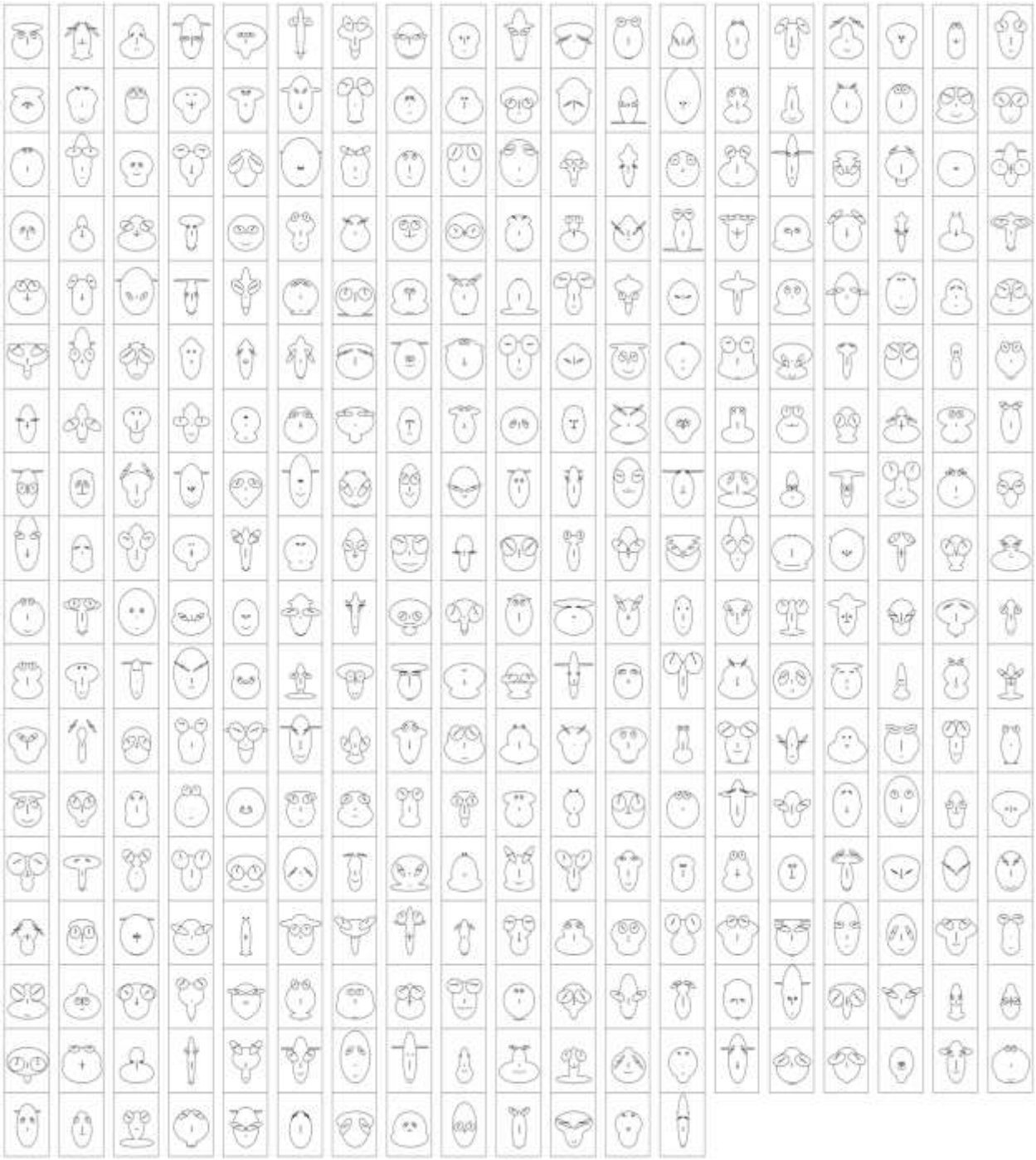

Figure 5.4: Chernoff Faces of All Records of One Output Label

5.4.3. IMAGE SLICING

Now the main and most important task is to slice the output image which is just one picture into multiples images. So for this purpose, we have to import the image_slicer library into python for slicing that image into multiple images according to the number of records. Furthermore, we

used the slice property of the image_slicer to slice the images. This property will detect the edged around each sub-image from the main image and then will crop this into a single image.

l

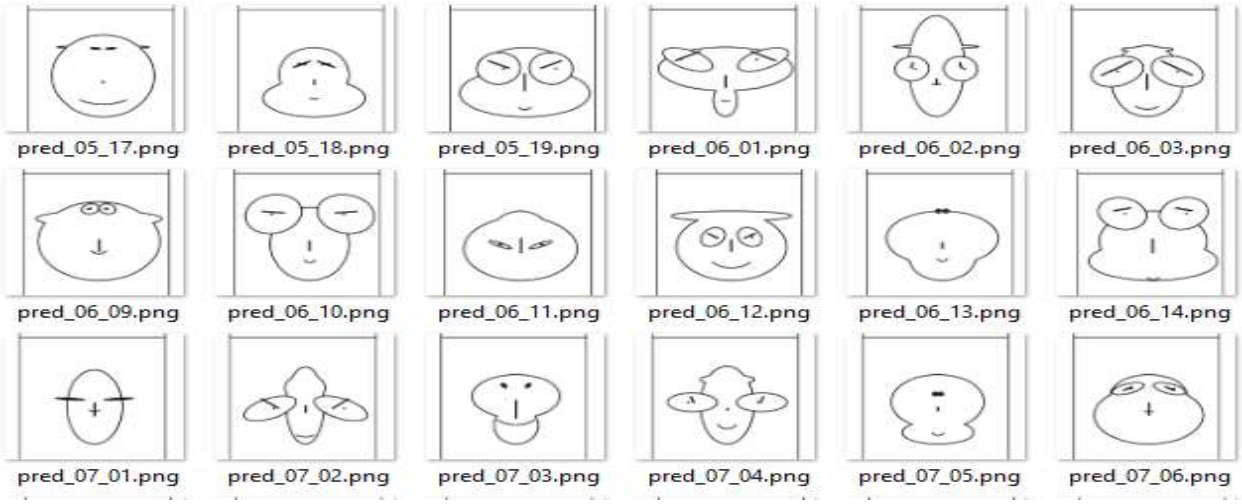

Figure 5.5: Image Slicing to Get Number of Images Equal to the Number of Records

Here is the output of the image_slicer that we got the sub-image of the main in the form of a single image. So we got the 336 images from the single image and got the same results for the rest of the three folders. After this now we can input these images into the CNN model for the extraction of the important features for the improvement of the accuracy of the dataset. We will provide the single image into the CNN model and extract features from that image and we will loop over the whole directory of the same output label.

5.5. FEATURE EXTRACTION USING CNN

Here in Figure 5.6 is the diagram of the Sequential CNN model which we developed for the features extraction and, we have done it excellently. We extracted the most important 27 features using this CNN model for classification and the improvement of accuracy of the dataset. The CNN model is very efficient in the process of feature extraction. And it has different properties like activation, padding, Strides, etc. These properties have done a great job of improving or increasing the effectiveness of the CNN model. This is the second last step of our proposed methodology after that, we will move towards the classification of the dataset which is then extracted from the Chernoff faces images.

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_1 (Conv2D)            (None, 181, 181, 8)       224
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 90, 90, 8)         0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 88, 88, 8)         584
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 44, 44, 8)         0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 42, 42, 8)         584
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 21, 21, 8)         0
_________________________________________________________________
conv2d_4 (Conv2D)            (None, 19, 19, 8)         584
_________________________________________________________________
max_pooling2d_4 (MaxPooling2 (None, 9, 9, 8)           0
_________________________________________________________________
conv2d_5 (Conv2D)            (None, 7, 7, 8)           584
_________________________________________________________________
max_pooling2d_5 (MaxPooling2 (None, 3, 3, 8)           0
_________________________________________________________________
dense_1 (Dense)              (None, 3, 3, 4)           36
_________________________________________________________________
dense_2 (Dense)              (None, 3, 3, 4)           20
_________________________________________________________________
dense_3 (Dense)              (None, 3, 3, 3)           15
_________________________________________________________________
flatten_1 (Flatten)          (None, 27)                0
=================================================================
```

Figure 5.6: Detailed Out Put of the CNN Model

Here is the detailed output of our Sequential CNN model. That explained the number of layers and their sequence. And also explain the height and width of the image and the kernel size and also defines the number of parameters for each convolution layer.

And Table 5.3 is the output detail below about the total parameters, trainable parameters, and non-trainable parameters. And in the end, we have the total number of images that were supplied to the CNN model as the input.

Table 5.3: Result of CNN Model

Total Parameters	2631
Trainable Parameters	2631
Non-trainable Parameters	0
Extracted Features	381

5.6. CLASSIFICATION RESULTS

For validation of our proposed methodology, we collect the accuracy results of three different datasets. Three different datasets are original data, selected features dataset, and extracted features dataset.

5.6.1. ORIGINAL DATASET RESULTS

This is the accuracy results of the four classifiers on the Original dataset after preprocessing in Table 5.4 given below.

Table 5.4: Accuracy results of four classifiers on Original dataset

Name of Dataset	Number of Features	Classifier	Accuracy Score
Original Dataset	29	Decision Tree	24.55%
		Random Forest	24.73%
		SVM	23.35%
		Bagging	24.54%

The Decision Tree classifier having a Score of 24.55% for 29 attributes. This is the original dataset which, we collected after the preprocessing of the data. This is free of anomalies and blunders. This score which we achieved is very low and we cannot trust the decision which will be taken based on that dataset. Some parameters, we used for the classification are test_size = '40'. And we used multi-classification for all the dataset forms. And here is the confusion matrix of the decision tree in Table 5.5 given below.

Table 5.5: Confusion Matrix of Decision Tree

Class	1	2	3	4
1	16	0	31	89
2	20	0	35	90
3	19	0	18	90
4	20	0	24	102

In Random forest classifier. The accuracy score we achieved is 24.73% with the test_size of 40 using a multi-class classifier. And here is the confusion matrix of the random forest classifier in Table 5.6 given below.

Table 5.6: Confusion Matrix of Random Forest

Class	1	2	3	4
1	1	0	104	31
2	2	0	113	30
3	4	1	95	27
4	5	0	100	40

In the Support Vector Machine (SVM) classifier and we achieved an accuracy score of 23.35%. Using a test_size of 40 for the multiclass dataset but still here not gaining good results for this dataset. But still, there is not too much good progress in achieving good results using Support Vector Machine. The obtained efficiency is very low for decision making. For proper decision making and to obtain good results we need to classify the data with high accuracy which would be reliable. And here is the confusion matrix of SVM in Table 5.7 given below.

Table 5.7: Confusion Matrix of SVM

Class	1	2	3	4
1	0	0	198	0
2	0	0	205	0
3	0	0	194	0
4	0	0	224	0

After the Support Vector Machine, we used bagging for the classification of the dataset having 29 attributes. Bagging is also good at classifying the data with a good accuracy score. We can see result in Table 5.8 given below.

Table 5.8: Confusion Matrix of Bagging

Class	1	2	3	4
1	0	0	198	0
2	0	0	205	0
3	0	0	194	0
4	0	0	224	0

We achieved an accuracy score of 24.54% using a bagging classifier. Using a multi-class classifier and testing size of 40%.

5.6.2. SELECTED ATTRIBUTES DATASET RESULTS

Here is the accuracy score of all four classifiers using the selected attributes dataset in Table 5.9 below.

Table 5.9: Accuracy score of four classifiers using selected attributes dataset

Data Name	Number of attributes	Classifier Name	Accuracy
Selected Features Dataset	17	Decision Tree	25.81%
		Random Forest	24.01%
		SVM	23.35%
		Bagging	22.56%

Selected attributes are those features which, we collected after applying different feature selection techniques like univariate selection, feature importance, and correlation matrix with heatmap. After applying these techniques we select the best 17 features and reduced those features which have low priority. After that, we apply different classifiers and collect the results of Decision Tree, Random Forest, SVM, and Bagging.

Table 5.10: Confusion Matrix of Decision Tree

Class	1	2	3	4
1	11	0	36	89
2	15	0	40	90
3	7	0	30	90
4	9	0	35	102

In Table 5.9 above the accuracy score of 25.81% using the decision tree classifier. We used different parameters like a test-size of 40 % and a multiclass classifier. The results are slightly better than the original dataset. And here is the confusion matrix of the decision tree in Table 5.10 given above.

Table 5.11: Confusion Matrix of Random Forest

Class	1	2	3	4
1	2	1	89	44
2	5	4	102	34
3	6	0	79	42
4	6	3	89	48

In Random Forest, we have an accuracy score of 24.01% using the Random Forest classifier for the selected feature dataset having a test-size of 40% and a multi-class classifier. In this case, the accuracy of the random forest classifier is decreased from the accuracy score of the original dataset. The accuracy score of the original dataset using the Random Forest classifier was 24.73% and the accuracy score of the Random Forest classifier for the selected features was 24.01%. This is lower than the first one. The decrease in the accuracy score of the Random Forest classifier could be because of the poor selection of the features of the high complexity of the data. But we select the best 17 features using three different approaches for the feature selection so the poor selection of the features could not be valid in this case. So the complexity

of the data or high correlation of different features with the output label can be the reason for low complexity. And here is the confusion matrix of random forest in Table 5.11 given above.

Table 5.12: Confusion Matrix of SVM

Class	1	2	3	4
1	0	0	198	0
2	0	0	215	0
3	0	0	194	0
4	0	0	224	0

Table 5.9 is the accuracy score of the Support Vector Machine for the 17 selected features. The accuracy of the SVM is also reduced from the accuracy of the original dataset using SVM. We used the sample test –the size of 40% and the Random state is equal to 42 and using a multi-class classifier. And here is the confusion matrix of SVM in Table 5.12 given above.

Table 5.13: Confusion Matrix of Bagging

Class	1	2	3	4
1	24	23	60	29
2	29	17	72	27
3	34	16	45	32
4	23	27	57	39

In the end, we have the bagging classifier for the classification of the seventeen selected features. The accuracy score, we achieved for bagging is 23.35% which is slightly less than the original dataset using the same classifier. The bagging classifier has done well in the classification of the original dataset than the selected features. We used the different parameters for the implementation of the Bagging classifier like sample test –the size of 40% and the Random state is equal to 42 and using a multi-class classifier. And here is the confusion matrix of bagging in Table 5.13 given below.

5.6.3. EXTRACTED FEATURES DATASET RESUL

Here is the output Table of different classifiers having different accuracy scores using the extracted features in Table 5.14 given below.

Table 5.14: Accuracy results of four classifier using extracted features

Dataset	Number of features	Classifier	Accuracy
Extracted Features	27	Decision Tree	97.29%
		Random Forest	94.58%
		SVM	97.40%
		Bagging	98.56%

Extracted features are those features that Are extracted from the Sequential CNN moel from Chernoff's faces. Chernoff's faces are built from those features which are selected using different selection techniques. Here is the accuracy score of the Decision Tree which is about 97.29%. The accuracy score which we obtained is very excellent for the decision making of the basis of that data. This accuracy score is very high from the accuracy score of the original dataset and the selected features dataset. This accuracy score is three times higher than the original and the selected feature datasets. And here is the confusion matrix of the decision tree in Table 5.15 given below.

Table 5.15: Confusion Matrix of Decision Tree

Class	1	2	3	4
1	129	0	0	0
2	0	134	8	0
3	0	7	119	0
4	0	0	0	157

Here is the accuracy score of the Random Forest in figure 5.19. The accuracy score we achieved from the Random Forest classifier is 94.58% which is also a very high number of accuracy for decision making. This accuracy score is lower than the result of the Decision Tree but still is very reliable for decision making. This is how our proposed methodology is performing in

increasing the accuracy of that dataset which is very complex and has very little accuracy using the same classifiers. Here we still using the Random_state= ‘42’ and the test-size is equal to 40% and using the same multi-class classifiers. And here is the confusion matrix of random forest in Table 5.16 given below.

Table 5.16: Confusion Matrix of Random Forest

Class	**1**	**2**	**3**	**4**
1	129	0	0	0
2	0	112	30	0
3	0	0	126	0
4	0	0	0	157

The output accuracy of the Support Vector Machine Classifiers results in the table above. The accuracy score we achieved using the SVM classifier is 97.47% which is very good for decision making. And this accuracy score is very high from the previous accuracy score of the original dataset and the seventeen selected features dataset. We are quite good at proving right out the proposed methodology. We used the same parameters for the implementation of the Support Vector Machine classifier. And here is the confusion matrix of SVM in Table 5.17 given below.

Table 5.17: Confusion Matrix of SVM

Class	**1**	**2**	**3**	**4**
1	80	0	0	0
2	0	73	9	0
3	0	0	89	0
4	0	0	0	95

In the end, we have the best accuracy score for the extracted features using a bagging classifier in the table above. The accuracy score we obtained using the bagging classifier is about 98.56% which is a very high level of accuracy for decision making. This is the high number of accuracy which we achieved using the extracted features using the Sequential CNN model. This is the highest accuracy in all three forms of data. And also three times higher than the accuracy of the original dataset and the selected features dataset. Using this accuracy rate we can rely on the decision making which will be taken based on that dataset. This is how our proposed

methodology improves the accuracy of the complex dataset from 24% to 99%. This is a very high level of margin to achieve. And here is the confusion matrix of bagging in Table 5.18.

Table 5.18: Confusion Matrix of Bagging

Class	1	2	3	4
1	129	0	0	0
2	0	135	7	0
3	0	1	125	0
4	0	0	0	157

The parameters used for the implementation of the bagging in python are the same as above and some additional features are n_estimators=10, max_samples=0.9, bootstrap=true, and n_jobs= -1. These parameters we used for the implementation of the bagging tree classifiers for the classification of the extracted features. And here is the combined table of the accuracy of all four classifiers using three different datasets in Table 5.19 given below.

Table 5.19: Results for All Three Dataset Using Different Approaches

	Dataset	No. of Features	Classifier	Accuracy Score
1	Original Dataset	28	Decision Tree	24.55%
			Random Forest	24.73%
			SVM	23.35%
			Bagging	23.47%
2	Selected Features	17	Decision Tree	25.81%
			Random Forest	24.01%
			SVM	23.35%
			Bagging	22.56%
3	Extracted Features	27	Decision Tree	97.29%
			Random Forest	94.58%
			SVM	97.40%
			Bagging	98.56%

And here are the best results from the previous researches given below in the Table 5.20 given below. We can see the difference between our and past results. The results that we achieved are much better than the previous researches. So our aim to achieve the improvement in the efficiency of supervised data via the CNN model was successfully achieved.

Table 5.20: Past Results using Python and R language

Year	**Author**	**Dataset**	**Technique /Lang**	**Accuracy**			
2020	Satish CR Nandipati et al.[17]	Hepatitis C virus (HCV) Egyptian	Python	Binary class Label		Multiclass Label	
				Number of Features	Accuracy	Number of Features	Accuracy
				12	50.39%	12	24.15%
				21	50.1%	21	24.44%
				29	50.15%	29	24.8%
			R	Binary class Label		Multiclass Label	
				Number of Features	Accuracy	Number of Features	Accuracy
				12	50.59%	12	50.03%
				21	49.57%	21	49.77%
				29	50.9%	29	50.12%

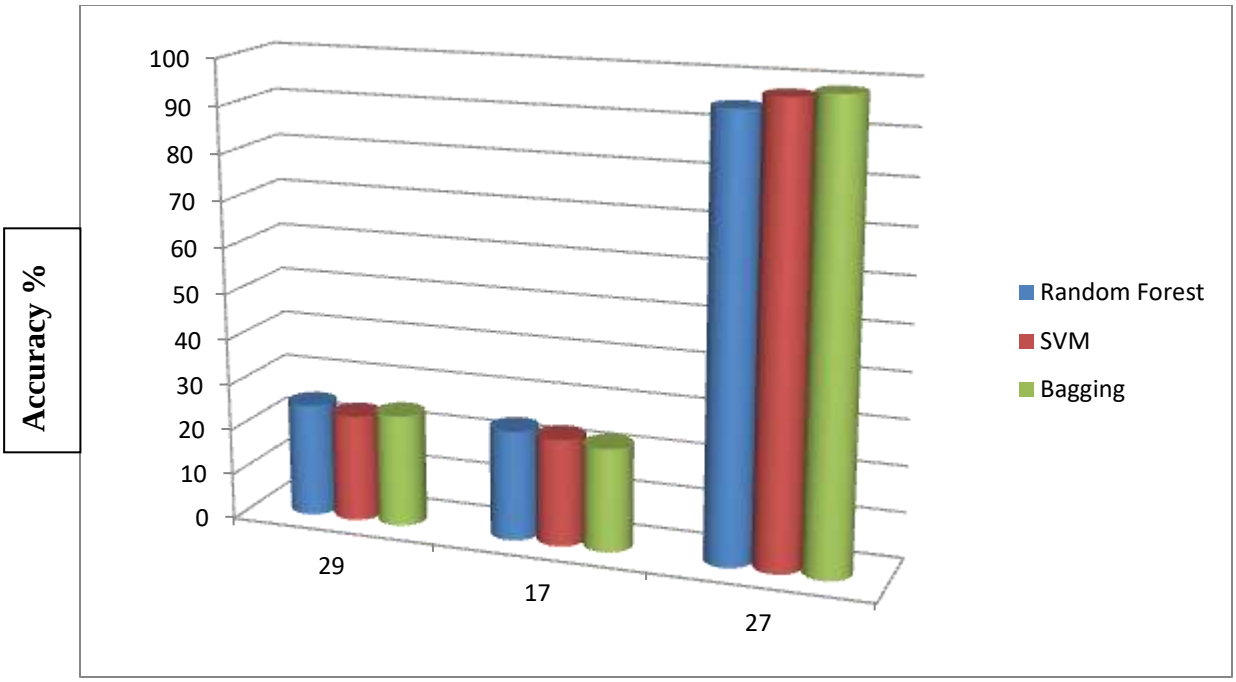

Figure 5.7: Our Proposed Methodology Results

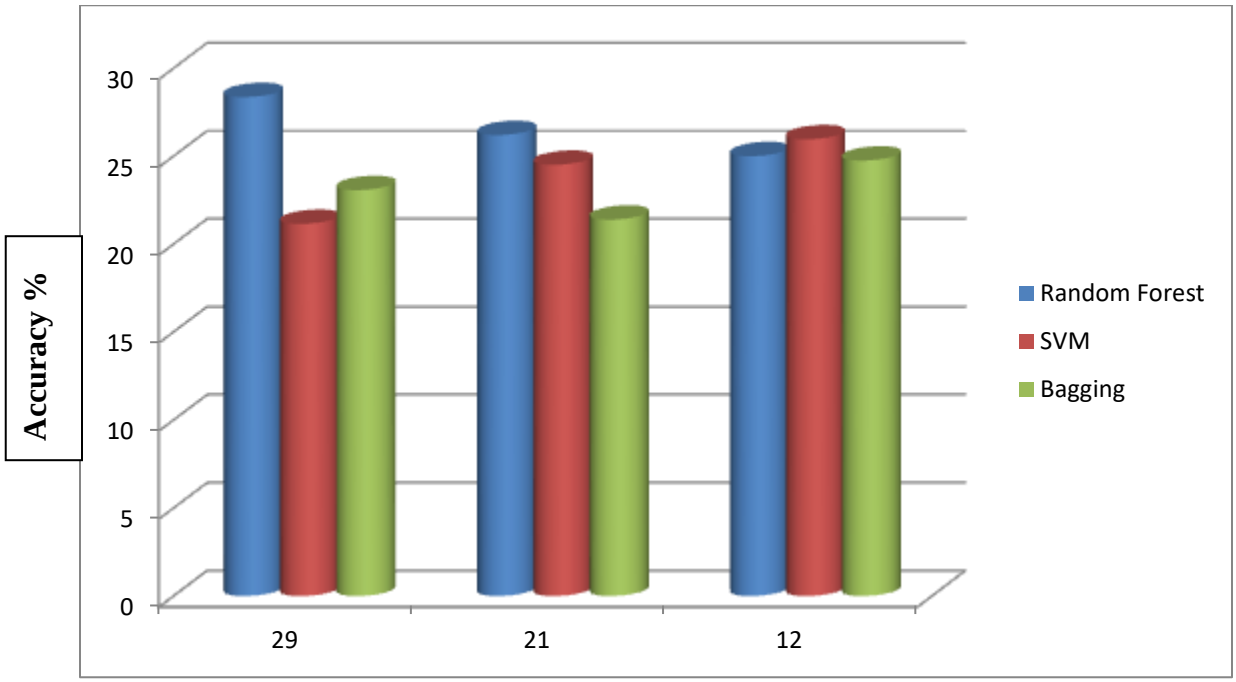

Figure 5.8: Results of past Researches

CHAPTER 6. CONCLUSION & FUTURE WORKs

6.1. CONCLUSION

The results show that the efficiency of the data is improved by using the proposed approach. Without using the approach the accuracy of prediction of the data was very poor by using a supervised data classifier. We used different classifiers to check the efficiency of the data like SVM, naïve Bayes, Decision tree, KNN, etc. But all of them show the poor accuracy of the dataset about 26% maximum by using a multiclass label and python. Some researchers also use binary class labels and R to improve the efficiency of the dataset and they have done this very efficiently. But our proposed methodology aimed to improve the efficiency of the dataset using the CNN model and we have done it excellently. After using the proposed methodology the results are different. As we use the Chernoff faces visualization technique for the numeric dataset which is the bone of the proposed methodology. And this visualization technique did the trick for us. And the remaining part of the proposed methodology increases the efficiency of the approach. And after using the proposed methodology we again use the same supervised classifier to check the accuracy of the data and this time the accuracy of the data jump to 99%. This is how our proposed methodology increases the efficiency of the Hepatitis C Virus (HCV) dataset.

6.2. CONTRIBUTION

- These pre-trained models could be used in better decision making using well manage data. Fraud Detection, investment plans, best product list for sale, and many more.

- This will increase the efficiency of data by differentiating between almost similar records using the CNN model a Chernoff's face.

- This will help us to do better decisions using data after this technique.

- Effective use of the memory

- Excellently deal with those datasets which have high dimensionality.

- More useful where the dataset has very low accuracy using supervised classifiers like SVM, KNN, and Naïve Bayes, etc.

- This will help the data analyst to have a better view of data having well-defined classes

6.3. FUTURE WORK

Our proposed methodology can be used for almost every supervised dataset with slight changes. It can increase the efficiency of the dataset. Faces are used for the visualizing of the numeric data and this will be the input for the CNN model. As we know CNN has different models like LeNet, AlexNet, VGG, GoogLeNet, ResNet, Sequential, and some variations of connections of layers as well as fully connected and have some dimensions variations all well like 1D, 2D, 3D CNN model. And all have their benefits. CNN model will automatically extract the feature from the images by using propagations. L2 regularization can play an important role in the determination of the accuracy for those datasets which have fewer numbers of training data. It will avoid the model by over-fitting. The number of features can be more or less by adding and subtracting numbers of layers accordingly. And then you can use those features for the classification of the data. For classification, we have many classifiers like Support Vector Machine (SVM), Naïve Bayes, Decision Tree, K-Nearest Neighbor (KNN), and many more. Our proposed methodology can be more useful for critical data like the medical and banking field. And also can be used for the investment plan.

REFERENCES

[1] A. Petruzziello, S. Marigliano, G. Loquercio, A. Cozzolino, and C. Cacciapuoti, "Global epidemiology of hepatitis C virus infection: An up-date of the distribution and circulation of hepatitis C virus genotypes," *World J. Gastroenterol.*, vol. 22, no. 34, p. 7824, 2016.

[2] Wikipedia, "Chernoff face - Wikipedia," 2020. https://en.wikipedia.org/wiki/Chernoff_face#References (accessed Oct. 28, 2020).

[3] Sandra Durcevic, "Data-Driven Decision Making – See 10 Tips For Your Business Success," 2016. https://www.datapine.com/blog/data-driven-decision-making-in-businesses/ (accessed Oct. 28, 2020).

[4] H. Chernoff, "The Use of Faces to Represent Points in K-Dimensional Space Graphically," *J. Am. Stat. Assoc.*, vol. 68, no. 342, p. 361, Jun. 1973, doi: 10.2307/2284077.

[5] T. Nocke, S. Schlechtweg, and H. Schumann, "Icon-based visualization using mosaic metaphors," in *Proceedings of the International Conference on Information Visualisation*, 2005, vol. 2005, pp. 103–109, doi: 10.1109/IV.2005.58.

[6] Wikipedia, "Stick figure - Wikipedia." https://en.wikipedia.org/wiki/Stick_figure (accessed Nov. 05, 2020).

[7] "Data Mining Definition." https://www.investopedia.com/terms/d/datamining.asp (accessed Nov. 05, 2020).

[8] internet, "Data Mining vs. Machine Learning: What's The Difference? | Import.io," 2017. https://www.import.io/post/data-mining-machine-learning-difference/ (accessed Nov. 05, 2020).

[9] "Data Mining - Classification & Prediction - Tutorialspoint." https://www.tutorialspoint.com/data_mining/dm_classification_prediction.htm (accessed Nov. 05, 2020).

[10] "Classification by Decision Tree Induction." http://www.brainkart.com/article/Classification-by-Decision-Tree-Induction_8321/ (accessed Jan. 23, 2021).

[11] H. Zhang, F. Wang, T. Xi, J. Zhao, L. Wang, and W. Gao, "A novel quality evaluation method for resistance spot welding based on the electrode displacement signal and the

Chernoff faces technique," *Mech. Syst. Signal Process.*, vol. 62, pp. 431–443, 2015.

[12] Y. Kim and L. Cooke, "Big data analysis of public library operations and services by using the Chernoff face method," *J. Doc.*, 2017.

[13] E. Gerhardt, "Visualization of Multi Key Performance Indicators by Dynamic Chernoff Faces."

[14] R. Song, Z. Zhao, and X. Wang, "An application of the V-system to the clustering of Chernoff faces," *Comput. Graph.*, vol. 34, no. 5, pp. 529–536, 2010.

[15] S. A. Moiz and R. R. Chillarige, "Method Level Code Smell: Chernoff Face Visualization," in *International Conference on E-Business and Telecommunications*, 2019, pp. 520–527.

[16] A. Haara, J. Pykäläinen, A. Tolvanen, and M. Kurttila, "Use of interactive data visualization in multi-objective forest planning," *J. Environ. Manage.*, vol. 210, pp. 71–86, 2018.

[17] S. C. R. Nandipati, C. XinYing, and K. K. Wah, "Hepatitis C Virus (HCV) Prediction by Machine Learning Techniques," *Appl. Model. Simul.*, vol. 4, pp. 89–100, 2020.

[18] S. M. Abd El-Salam *et al.*, "Performance of machine learning approaches on the prediction of esophageal varices for Egyptian chronic hepatitis C patients," *Informatics Med. Unlocked*, vol. 17, p. 100267, 2019.

[19] S. Hashem, G. Esmat, W. Elakel, … S. H.-I. T., and undefined 2018, "Comparison of Machine Learning Approaches for Prediction of Advanced Liver Fibrosis in Chronic Hepatitis C Patients," *dl.acm.org*, Accessed: Nov. 06, 2020. [Online]. Available: https://dl.acm.org/doi/abs/10.1109/TCBB.2017.2690848.

[20] S. Hashem, G. Esmat, W. Elakel, S. Habashy, and S. Raouf, "Comparison of Machine Learning Approaches for Prediction of Advanced Liver Fibrosis in Chronic Hepatitis C Patients," *gamalesmat.com*, Accessed: Nov. 06, 2020. [Online]. Available: http://www.gamalesmat.com/PDFs/171.pdf.

[21] S. Hashem, G. Esmat, W. Elakel, … S. H.-G., and undefined 2016, "Accurate Prediction of Advanced Liver Fibrosis Using the Decision Tree Learning Algorithm in Chronic Hepatitis C Egyptian Patients," *ncbi.nlm.nih.gov*, Accessed: Nov. 06, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4736594/.

[22] S. Hashem *et al.*, "Accurate prediction of advanced liver fibrosis using the decision tree

learning algorithm in chronic hepatitis C Egyptian patients," *Gastroenterol. Res. Pract.*, vol. 2016, 2016.

[23] S. Hashem, G. Esmat, W. Elakel, … S. H.-G., and undefined 2016, "Accurate Prediction of Advanced Liver Fibrosis Using the Decision Tree Learning Algorithm in Chronic Hepatitis C Egyptian Patients," *airitilibrary.com*, Accessed: Nov. 06, 2020. [Online]. Available: https://www.airitilibrary.com/Publication/alDetailedMesh?docid=P20161005001-201612-201711070015-201711070015-432-438-056.

[24] S. Hashem, G. Esmat, … W. E.-G., and undefined 2016, "Accurate Prediction of Advanced Liver Fibrosis Using the Decision Tree Learning Algorithm in Chronic Hepatitis C Egyptian Patients," *search.proquest.com*, Accessed: Nov. 06, 2020. [Online]. Available: http://search.proquest.com/openview/d05a307eacf7a25b2b886ed097899af9/1?pq-origsite=gscholar&cbl=2037462&casa_token=LnnwMDcKRnUAAAAA:bXgfwrs72qtxwIHL4UY3bkQYCsQvdSoYCoEPasjxGysaTyO8M1Afw77bTnLytIXVYQgP_-YUHCk.

[25] S. Hashem, G. Esmat, W. Elakel, S. Habashy, and S. Raouf, "Accurate Prediction of Advanced Liver Fibrosis Using the Decision Tree Learning Algorithm in Chronic Hepatitis C Egyptian Patients," 2016, Accessed: Nov. 06, 2020. [Online]. Available: http://central-library.msa.edu.eg:8009/xmlui/handle/123456789/1444.

[26] S. Hashem, G. Esmat, W. Elakel, S. Habashy, and S. Raouf, "Accurate Prediction of Advanced Liver Fibrosis Using the Decision Tree Learning Algorithm in Chronic Hepatitis C Egyptian Patients," 2016, Accessed: Nov. 06, 2020. [Online]. Available: http://www.academia.edu/download/44000754/Accurate_prediction_of_fibrosis_by_Model.pdf.

[27] S. Hashem, G. Esmat, W. Elakel, S. Habashy, and S. Raouf, "Accurate Prediction of Advanced Liver Fibrosis Using the Decision Tree Learning Algorithm in Chronic Hepatitis C Egyptian Patients," 2016, Accessed: Nov. 06, 2020. [Online]. Available: http://www.gamalesmat.com/Admin/uploads/Accurate_Prediction_of_Advanced_Liver_Fibrosis_Using_the_Decision_Tree_Learning_Algorithm_in_Chronic_Hepatitis_C_Egyptian_Patients.pdf.

[28] G. G. Agarwal, A. K. Singh, V. Venkatesh, and N. Wal, "Determination of risk factors for

hepatitis C by the method of random forest," *Ann. Infect. Dis. Epidemiol.*, vol. 4, no. 1, 2019.

[29] S. Hashem *et al.*, "Comparison of Machine Learning Approaches for Prediction of Advanced Liver Fibrosis in Chronic Hepatitis C Patients," *gamalesmat.com*, vol. 15, no. 3, pp. 861–868, May 2016, doi: 10.1109/TCBB.2017.2690848.

[30] S. Hashem *et al.*, "Comparison of machine learning approaches for prediction of advanced liver fibrosis in chronic hepatitis C patients," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 15, no. 3, pp. 861–868, 2017.

[31] S. Hashem, G. Esmat, W. Elakel, … S. H.-… /ACM transactions on, and undefined 2018, "Comparison of Machine Learning Approaches for Prediction of Advanced Liver Fibrosis in Chronic Hepatitis C Patients.," *europepmc.org*, Accessed: Nov. 06, 2020. [Online]. Available: https://europepmc.org/article/med/28391204.

[32] S. Hashem, G. Esmat, W. Elakel, … S. H.-I. A. of the, and undefined 2018, "Comparison of Machine Learning Approaches for Prediction of Advanced Liver Fibrosis in Chronic Hepatitis C Patients," *computer.org*, Accessed: Nov. 06, 2020. [Online]. Available: https://www.computer.org/csdl/journal/tb/2018/03/07891989/13rRUxASufU.

[33] S. Hashem, G. Esmat, … W. E.-… A. transactions on, and undefined 2017, "Comparison of machine learning approaches for prediction of advanced liver fibrosis in chronic hepatitis C patients," *ieeexplore.ieee.org*, Accessed: Nov. 06, 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7891989/.

[34] B. Liu, X. Yu, P. Zhang, A. Yu, Q. Fu, and X. Wei, "Supervised deep feature extraction for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 1909–1921, 2017.

[35] A. Yang, X. Yang, W. Wu, H. Liu, and Y. Zhuansun, "Research on feature extraction of tumor image based on convolutional neural network," *IEEE Access*, vol. 7, pp. 24204–24213, 2019.

[36] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, 2016.

[37] "Hepatitis C Drugs Not Reaching the Poor - Scientific American." https://www.scientificamerican.com/article/hepatitis-c-drugs-not-reaching-the-poor/

(accessed Jan. 23, 2021).

[38] “Chernoff Faces Examples | Evolving world melodies one tune at a time.” http://horvathcampbell.com/chernoff-faces-2/ (accessed Jan. 23, 2021).

[39] “Faces 2.2 - bradandkathy.com.” https://bradandkathy.com/software/faces.html (accessed Jan. 23, 2021).

[40] “Decision Tree Classification. A Decision Tree is a simple… | by Afroz Chakure | The Startup | Medium.” https://medium.com/swlh/decision-tree-classification-de64fc4d5aac (accessed Jan. 23, 2021).

[41] “Random Forests.” https://kevintshoemaker.github.io/NRES-746/RandomForests.html (accessed Jan. 23, 2021).

[42] W. Blake *et al.*, “Accurate Prediction of Advanced Liver Fibrosis Using Decision Tree Accurate Prediction of Advanced Liver Fibrosis Using the Decision Tree Learning Algorithm in Chronic Hepatitis C Egyptian Patients,” 2016, doi: 10.1155/2016/2636390.

[43] S. Hashem, G. Esmat, W. Elakel, and S. H.-… research and practice, “Accurate prediction of advanced liver fibrosis using the decision tree learning algorithm in chronic hepatitis C Egyptian patients,” *hindawi.com*, Accessed: Nov. 06, 2020. [Online]. Available: https://www.hindawi.com/journals/grp/2016/2636390/abs/.

%