ISSN 2455 – 9091

IJCSET JANUARY Volume 9 Issue 1

International Journal of Computer Science Engineering and Technology (IJCSET)

https://www.doi.org/10.5281/zenodo.8401859

Depression Analysis Using Supervised Models

[1] G. Ganapathy Babu
[1]Assistant Professor,
[1] Department of Computer Science Engineering,
[1] St. Martins Engineering College, Secunderabad, Telangana.

ABSTRACT: Depression is one of the severe and gravehealthdisordersthataffectsthesteadinessofmind.Ithasbecomeaseriousissueinthepresentgeneration.Thetotalnumber of cases has been increasing day-by-day due to anumber of reasons like stress at school, college, work,personal life, other diseases, etc. Although it has becomeoneofthemostcommondisease,peoplearestillreluctantto talk about it openly due to the fear that others mightconsiderthemlunatic.TheintroductionofMachineLearning into the field of Medicine and Health industryhasprovideddiagnostictoolsthatareabletoenhanceth eprecision and accuracy while reducing the difficult taskswhichrequiretheinterventionofhumans.Thereispromisingevidencethat MachineLearninghasthecapabilitynotonlytodetectbutalsosignificantlyenhancethe treatment of compound mental conditions such asdepressionbydevelopingaframework.Inthepast,Machine Learning Algorithms have been proved to befairly supportive where researchers worked on the datafromsocialmediatoforeseethenumberofpersonssuffering from this ailment on the basis of their initialsymptoms. The main aim is to help those patients whosufferfromdepressionintheearlyrecognitionofsymptoms of this disease which can prove to be valuablenotonlytothembutalsototheirfamilies.

Keywords: [Depression, Machine Learning, Naïve BayesAlgorithm,SupportVectorMachine,DecisionTree,Random Forest, KNearest NeighbourClassifier.]

1. INTRODUCTION

Depressionisconsideredtobeaverylethalanddistinctive medical sickness whichis relatedtothepsychological health of an individual. It is also knownby another name, Major Depressive Disorder (MDD).It is a disorder that affects a large number of peopleacrosstheglobe,around264millionpeopleareaffected by depression worldwide. The characteristicsof this disorder are: losing concentration in severalactivities,extrememoodswings,sleepingillnesses,demolishingsadnesswhichisthenfollowedbyreluctance in getting up in the morning, exhaustion,disinterested in the daily routine, abundant crying, etc.Depression can seriously affect the day-to-day life of aperson from the least possible work to any major task.These days people even consider depression as a tabooandchoosenottodiscussaboutitwithanyoneorin fact,theyevenholdbackfromconsideringthisadisease or a grave disorder that is required to be cured.Asaconsequenceofthis,peoplewhoaregoingthroughthis condition from are fearful to even talk about it. Thisnotonlymakesthemsufferquietly,withoutgettinganysort of medical help but also worsens their conditionfurther [1].

Depression has become one of the grave conditions thathas triggered the suffering among numerous peopleworldwide. This has become one of the leading causesof worsening in their day-to-day life. Early discoveryofthisillnessisveryimportantasitcanhelpanindividualtogetthedesiredmedicalaidintime,whichcan help him/her lead a normal healthy life and makehis/herconditionbetterinsteadofmakingitworse.Theobjectiveisgivenas:

Perceiving in advance the chances of a person tosuffer from depression by asking him/her certainstandard questions.

Findingthemostaccuratemachinelearningalgorithm by comparing several algorithms andchecking theiraccuracy.

Knowingthedepressions'severitylevelforaparticularperson.

2. LITERATURESURVEY

To appreciate the role of machine learning in mentalillnessdiagnosis,anumberofpapersandjournalshaveb een presented. The study on this matter started in1980's. Roland H.C. Yap, David M. Clarke portrayedintheirpaperaskilledsystem,MILP(MonashInterviewforLiaisonPsychiatry)thatcanidentifymental condition based on DSM-III-R, DSM-IV andICD-10 using constraint-based reasoning. ConstraintLogicProgramming(CLP)languagewasusedtodevelop thissystem[2].

There is a dataset from Reddit that was published in2017 which is publicly accessible.It was taken fromRedditInc.API,thatencompassedfeaturessuchas:id,title,writing,date.Typically,itcomprisedofreports thatwereprovidedbypersonsdiagnosedwithdepression.Itwasconsideredtobeacharacterizeddataof depression. A control group also shared their datathat was considered to be non-depressed characterizeddata. A total of 135 subjects were found to be depressedoutof887total subjects[3].

Thereisanotherdatasetwherethedatafromadepression survey was joined with the already presentpublicsocialmediasources.Thesourcesincludedself-declareddepressioncases,differentforums,self-reportedsurveysand postlevel annotation[4].

There is another publicly available dataset that is used.ItisfromOpenSourcingMentalIllness(OSMI)surveyth at was conducted in 2017. The survey was mentalhealth tech survey which consisted of a total of 750responseswith68attributes[5].

One of the common practices in today's world is theuseemojisandslangs onsocialmediasiteswhichmakesitdifficulttointerpret.Toresolvethisissue,bothslangs and emojis were replaced with their descriptivetextbymakingtheuseofanSMSdictionaryandemojipedia. To reduce the words to their respective rootwords,stemmingwasdone andforthe purposeoffiltering out various tokens, tokenization was done [6].Some researchers did the analysis of depression on theFacebookdatawhichisavailablepublicly.Theycarried out their research on the basis of linguistic andemotional style of word usage [7]. The classificationwas done by making the use of SVM algorithm withdifferent kernels and the researchers showed that theSVMalgorithmoutperformedwithbetteraccuracy[8].One oftheresearchersnamedNadeemconductedexperimentin2016onMajorDepressiveDisorder(MDD) via twitter data. He made the use of NaïveBayes and SVM algorithm. His research concluded with Naïve Bayesalgorithmoutperforming SV Malgorithm [9]. Inoneofthestudies,ahybridmodelofmachinelearning was implemented on the twitter data for thedetectionofdepression.TheSVM-NaïveBayeshybridmodelshowedagreataccuracyforthetaskofsentimentanalysis[10].

3. PROPOSEDSYSTEM

Theexistingmodelsthathavebeencreatedhaveusually worked on the datasets that were based on thesocial media like twitter and Facebook. These datasetsarenotusuallyreliableaspeopletendtobedifferentonsocial media then what they actually are in real life. Theposts on the social media sites are usually copied fromanother person's account or from some other source.So,aproperdiagnosiscannotbemadeonsuchasituation.Therefore,adifferentapproachhasbeen taken in this work where a person is asked a series ofstandard questions and based on his responses to thesequestions,apredictionismadebythemodelwhichhasbeentrainedusingadatasetwhichcontainsabout16000entries.FiveMachineLearningAlgorithmswereusedtotrainthe model:
SVMClassifier
DecisionTreeClassifier
RandomForestClassifier
NaïveBayesClassifier
KNNClassifier
SVMClassifierachievedthebestaccuracyandthuswasimplementedinthe application.

1. Support Vector Machine (SVM)Classifier

SVMisthemostprevalentSupervisedLearningalgorithms. Owing to its simplicity, it surpasses all theotheralgorithmswhenitcomestoitsusage[1].Itfindsits use in both Classification and Regression purposes.However, in Machine Learning, it is mainly used forClassification problems. The objective of SVM is tocreateadecisionboundarythatcanseparaten-dimensional space into classes so that a new data pointcan be inserted in the correct category in the future. Thebest decision boundary created by SVM is called thehyperplane. The algorithm is called Support VectorMachineasitchoosestwoextremepoints/vectorscalled support vectors which help in the creation of thehyperplane [11].

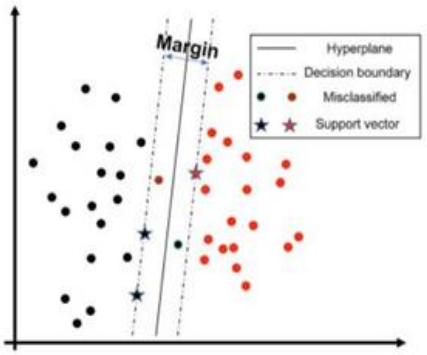

Figure. 1. SVM Classifier

[12]. SVM can beoftwotypes:

Linear SVM: It is used for Linearly Separable data,which means if a dataset can be classified into twoclassesbyusingasingle straightline
Non-linearSVM:Itisusedfornon-linearlyseparabledata,whichmeansadatasetthatcannotbeseparatedbyusing a single straightline.

DecisionTree Classifier

It is a Supervised Learning classifier which is preferably used to solve classification problems. Asits namesuggests, Decision Tree has the structure of a tree in which the branches signify the rules of decision, internal nodes signify the features of the dataset andeachleafnodecharacterizestheoutcomeofadecision.
Thereare two types of nodes in the decisiontree, Decisionnodeandleafnode. Decisionnodes havemultiple branches and are used to make any decision.Leaf nodes do not have any branches as they are theoutcomeof thedecisionofdecisionnodes.

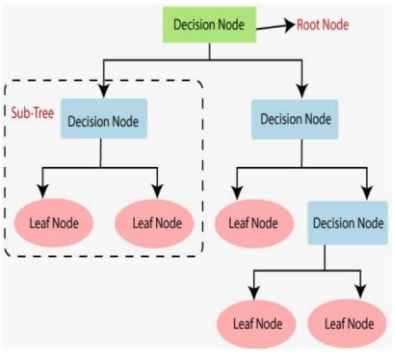

Figure. 2. Decision Tree Classifier [13].

The working of a decision tree is similar to a normaltree which starts with the root node and then expandsfurtherbybranchingitselfandformingatree-likestructure.Inordertobuildthetree,the algorithmwhichwe used is called Classification and Regression Tree(CART) algorithm. In simple terms, a decision treeasks a question and based on the answer which is YesorNo, itsplitsfurtherintothesubtrees.

Random Forest Classifier

RandomForestClassifierisaprevalentmachinelearning algorithm which comes under the supervisedlearningtechnique.InML,itisusedforbothClassification and Regression problems. The basis forthis algorithm is the notion of ensemble learning, whichcan be defined as a process of combining numerousclassifiers to

solve a composite problem and also toimprovethe performanceofthe model.

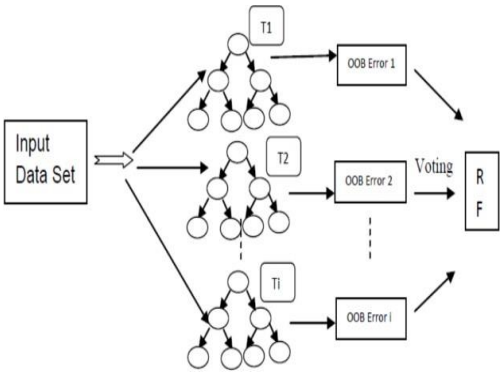

Figure. 3. Random Forest Classifier [14].

A Random Forest algorithm produces several decisiontreesinplaceofasingletree.Whenanewinputfromagivensampleisfedtorandomforestforclassification purpose, each of its trees is given that same input toperformtheclassification.Theclassificationfromeachtreeis called"votes"forclassifiedclass.Theclassificationwhichreceivesmaximumnumberofvotesisselected[14].Themorethenumberofdecisiontreesintherandomforest,thehigheristheaccuracyofthe algorithm and also the problem of overfitting isreduced.Thereisapossibilitythatsomeofthedecisiontrees in the forest may predict the incorrect output. Butsince there are a number of trees in the forest, othertrees predict the correct output and hence the finaloutputpredictediscorrect.

Naïve Bayes Classifier

It is a supervised algorithm which is based on Bayestheorem. It is used for classification problems mainlyintextclassificationwhichincludesahigh-dimensionaltrainingdataset.Itisoneofthemostsimple and effective classification algorithms that helpsus to build fast Machine Learning models that can makequick predictions. Since this classifier predicts on thebasisoftheprobability,itiscalledprobabilisticclassifier. It is composed of two words: Naïve andBayeswhichare describedas [15]:

Naïve: as it assumes that a certain feature's occurrenceisindependentoftheoccurrence ofotherfeatures.

Bayes:asitdependsontheprincipleofBayes'Theorem.

TheNaïveBayestheorem isshowninfig.4.

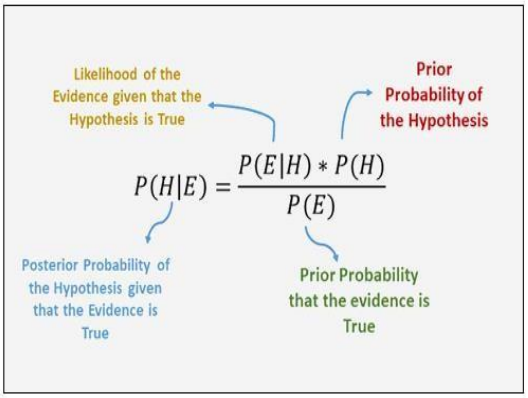

Figure.4. Naïve Bayes Classifier [16]

In the field of Probability and statistics Bayes theoremis considered to be vital. It defines the probability of theoccurrence of an event based on some conditions thatareknownprior totheuser.

KNNClassifier

K-NearestNeighbourisasupervisedlearningtechnique-basedalgorithm.Itisanalgorithmthatworksontheassumptionthatthereisasimilaritybetween the new data and the available data. A newcase/dataisputinto thecategorywhich ismostsimilar

to the available categories. k-NN stores all the existingdata and a new data point is categorized based on thesimilarity which means that when a new data appears,it can be classified into a well suite category by usingk-NN algorithm. This algorithm is mainly used forclassification but it can also be used for regressionpurposes.Itisanon-parametricalgorithmthatdoesnotmake any assumption on the original data. Since thisalgorithmdoesnotimmediatelylearnfromthetrainingdata set, it is sometimes referred to as lazy learneralgorithm. In the training phase it just only stores thedataset and when it gets the new data then it performsthe classification on the data and classifies it into acategory that ismuchaliketothenew data.

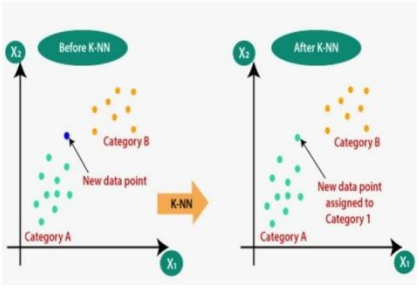

Figure.5. KNN Classifier [17]

A number of data pre-processing steps are performedonthecsvfilethatisread.Naturallanguageprocessinghas been used for pre-processing methods applied ontheextracteddata:

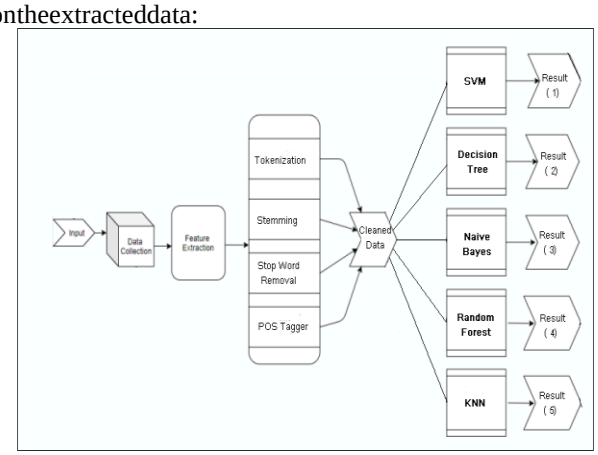

Figure.6. Workflow Model Architecture.

Tokenization:

Dividingastringintoseveralmeaningfulsubstringslikewords and sentences.

Stemming: It involves reducing the words to the irroot form so astogroup similarwordstogether.

StopWordRemoval:Stopwordslikea,an,theetcneed tobe removed since theyare ofnouse.

POS Tagger: Tokenized words are assigned tagssuch as nouns, adjectives etc to improve the qualityof thetraineddata.

EXPERIMENTALRESULTS

We trained all the five models and upon testing eachproduced a different result. The interpretation of theresultswasdonebymakingtheuseofconfusionmatrix(figure7-11).The labelscanbeinterpretedas:

0 - No Depression1-MildDepression
–ModerateDepression
– Moderately Severe Depression4 –Severe Depression

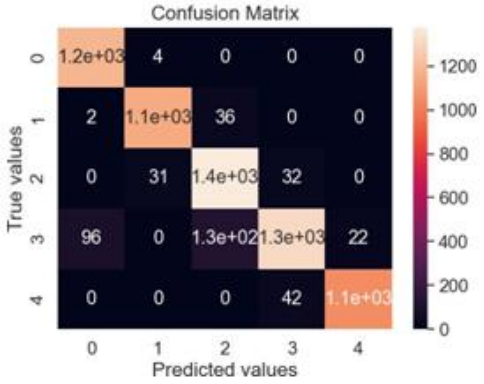

Figure.7. SVM Classifier.

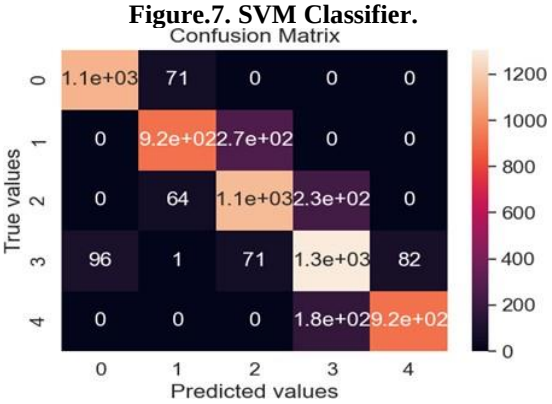

Figure.8. Random Forest Classifier.

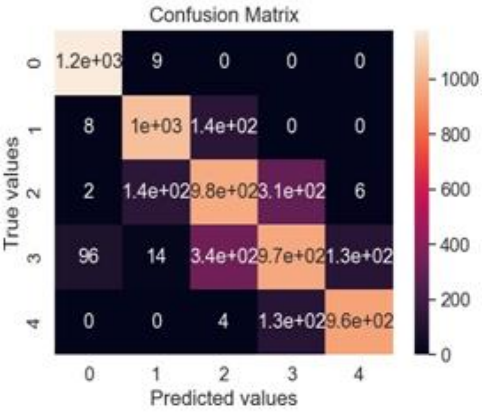

Figure.9. DT Classifier.

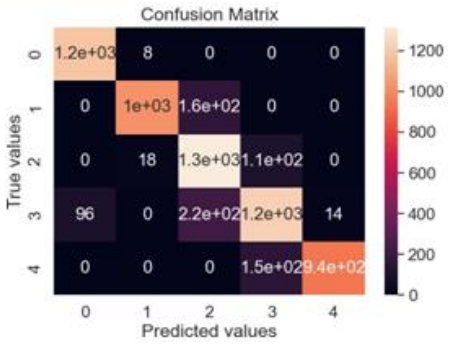

Figure.10.NaïveBayesClassifier.

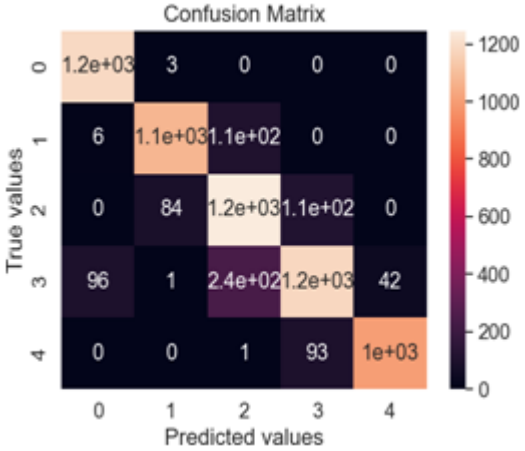

Figure.11.KNNClassifier.

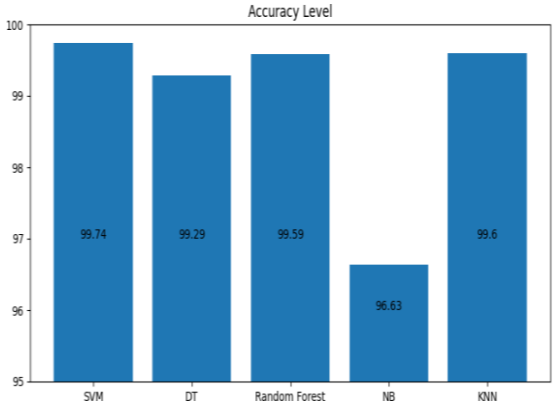

Figure. 12. GraphicalRepresentationofresults.

The results obtained by different algorithms are givenin table 1 and its graphical representation is given inFig. 12.

ALGORITHMSUSED	ACCURACY
SUPPORTVECTORMACHINE(SVM)	99.74
DECISIONTREECLASSIFIER	99.29
RANDOMFORESTCLASSIFIER	99.59
NAÏVEBAYESCLASSIFIER	96.63
KNNCLASSIFIERCLASSIFIER	99.60

Table 1. Accuracies obtained from various MachineLearningalgorithms.

CONCLUSION

Since depression is a mental health ailment which isspread across the world, immediate dealing with it hasbecome necessary. Some of the most critical steps todeal with this disorder are an early detection of thisdiseasethroughcommonsymptomsandmorewidespread knowledge about it. These two steps canaid the people in getting better treatment and can alsosave many lives. This work was initiated with the goalofhelpingthepeopleinneedbypredictingthedepression in its early stages so that they can get cureintime.Inthiswork,differentmachinelearningalgorithms wereused,andalsovariousfeaturedatasetstotrainthemodel.Preparationofdataanditsalignment, labelling of data and feature extraction andselection are a few of the pre-processing procedures.There are some of the other existing systems that arebased on numerous other techniques like - SupportVectorMachineswithanaccuracyof85.71%,Boosting which gives an accuracy of 75%, ConvolutionNeural Network-based prototype with diverse

featuresfedintothemodelwithanaccuracyof95%,andanother model which uses random forest, having anaccuracyof81.04%.Dataanalysiswasdonethoroughly to define the participants' behaviour basedonmanyfeaturesoftheirPHQ9questionresponses.Onth e basis of the conclusions drawn from the trainedmodel, the result of the research done was divided intofive labels according to the severity of the depression.Different machine learning algorithms that have beenusedinthispaperareasfollows:Naïve-Bayestheorem,Support Vector, k-Nearest Neighbour, Decision TreeandRandomForest.Theaccuraciesachievedinascending orderareNaiveBayesclassifierwith96.6%,Decision Tree classifier with 99.2%, Random Forestwith99.5%,KNNclassifierwith99.6%andSVMwith9 9.7%.

FUTUREENHANCEMENTS

Depression is a perplexing mental health disorder i.e.;itisveryhardtounderstandeverythingaboutit.Sometimest hesymptomsareverycommonandconfined only to some health issues that are very basicbutothertimestheymaybeveryobvioustotheperson.This limit sometimes makes it really hard for the personto diagnose and get appropriate treatment. There is acompromise in both the excellence as well as the sizeof the dataset. As a consequence of this, usually thework has to be done on a small dataset. In the futureany work done on this subject should be done on adataset that is both large with a greater number ofattributesandalsothequalityshouldbebettersothatit may be trustworthy and can achieve a more promisingresult. Neural network-based models can also be builtas an improvement to the present work to check theirperformanceandprecision.

REFERENCE

[1]. G.Geetha, G.Saranya, Dr.K.Chakrapani, Dr.J.Godwin Ponsam, M.Safa, Dr.S.Karpagaselvi. "EARLYDETECTIONOFDEPRESSIONFROMSOCIAL MEDIADATAUSINGMACHINELEARNING ALGORITHMS "20202nd International Conferenceon Power, Energy, Control and Transmission Systems

[2]. RH Yap and David M Clarke. "An expert systemfor psychiatric diagnosis using the dsm-iii-r, dsm-iv and icd-10 classifications." In Proceedings oftheAMIAAnnualFallSymposium,page229.AmericanMe dicalInformaticsAssociation,1996.

[3]. Fidel Cacheda, PhD;Diego Fernandez, PhD; FranciscoJNovoa1,PhD;VictorCarneiro,PhD,"EarlyDetect ionofDepression:SocialNetworkAnalysisandRandomFores tTechniques",2019.

[4]. SharathChandraGuntuku,DavidBYaden,Margaret L Kern, Lyle H Ungar and Johannes CEichstaedt,"Detectingdepressionandmentalillness on social media: an integrative review",2017.

[5]. Kali Cornn, Department of Statistics, Stanford University, "Identifying Depressionon SocialMedia",2018.

[6]. AkshiKumar, Aditi Sharma, Anshika Arora," Anxious Depression Predictionin Real-timeSocialData".2019.

[7]. M.R.Islam,A.R.M.Kamal,N.Sultana,R.Islam, M.A.Moni,andA.Ulhaq,"DetectingDepressionUsing K-Nearest Neighbors (KNN) ClassificationTechnique," Int. Conf. Comput. Commun. Chem.Mater.Electron.Eng.IC4ME22018,pp.4–7,2018,doi:10.1109/IC4ME2.2018.8465641.

[8]. H.S.ALSAGRIandM.YKHLEF,"MachineLearning-BasedApproachforDepressionDetection in Twitter Using Content and ActivityFeatures," IEICE Trans. Inf. Syst., vol. E103.D,no.8,pp.1825–1832,2020,doi:10.1587/transinf.2020edp7023.

[9]. M. Nadeem, "Identifying Depression on Twitter,"pp. 1–9,2016.

[10]. M. Gaikar, J. Chavan, K. Indore, and R. Shedge,"Depression Detection and Prevention System byAnalysing Tweets," SSRN Electron. J., pp. 1–6,2019,doi:10.2139/ssrn.3358809.

[11]. https:// www.javatpoint.com/ machine-learningsupport- vector- machine- algorithm

[12]. Siddharth Misra, Hao Li, "Non-invasive fracturecharacterizationbasedontheclassificationofsonic wave travel times" Machine Learning forSubsurface Characterization,2020.

[13]. Muhammad Asfand Hafeez, Muhammad Rashid,Hassan Tariq, Zain Ul Abideen, Saud S. AlotaibiandMohammedH.Sinky"PerformanceImprovemen tofDecisionTree:ARobustClassifierUsingTabu SearchAlgorithm"

[14]. Veena N. Jokhakar and S.V. Patel "A RandomForestBasedMachineLearningApproachforMildSt eelDefectDiagnosis"2016IEEEInternationalConferenceon ComputationalIntelligenceandComputingResearch (ICCIC)

[15]. https://www.javatpoint.com/machine-learning-naive-bayes-classifier

[16]. https://medium.com/analytics-vidhya/na%C3%AFve-bayes-algorithm-5bf31e9032a2

[17]. https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning.