2023-11

DEEP LEARNING-BASED BUSINESS INCOME TAX FRAUDDETECTIONMODE

ANEGAW, SISAY TESFAYE

BAHIRDARUNIVERSITY

BAHIRDARINSTITUTEOFTECHNOLOGYS

CHOOL OFGRADUATESTUDIES

FACULTY OF COMPUTING

DEPARTMENTOFINFORMATIONTECHNOLOGY

MSc THESIS ON:-

DEEP LEARNING-BASED BUSINESS INCOME TAX FRAUDDETECTIONMODEL

BY:

ANEGAWSISAYTESFAYE

NOVEMBER,2023

BAHIRDAR,ETHIOPIA

BAHIRDARUNIVERSITY

BAHIR DAR INSTITUTE OF

TECHNOLOGYFACULTYOFCOMPUTING

DEEPLEARNING-BASEDBUSINESSINCOMETAXFRAUDDETECTIONMODEL

By:

AnegawSisayTesfaye

AThesisSubmittedToBahirDarUniversity,BahirDarInstituteOfTechnology,SchoolOf Graduate Studies. In PartialFulfillmentOf The Requirements ForThe DegreeOf MasterofScienceInTheInformationTechnologyInTheFacultyOfComputing.

Advisor:-Dr. MekonnenWagaw (PhD)

NOVEMBER,2023

BAHIRDAR,ETHIOPIA

DECLARATION

I, the undersigned, declare that the thesis comprises my own work. In compliance with Internationally accepted practices, I have acknowledged and refereed all materials used in this Work. I understand that non-adherence to the principles of academic honesty and integrity, Misrepresentation/ fabrication of any idea/data/fact/source will constitute sufficient ground for disciplinary action by the University and can also evoke penal action from the sources which have not been properly cited or acknowledged?

AnegawSisayTesfaye	________________	NOVEMBER,2023
Nameofthecandidate	signature	Date

APPROVAL SHEET

BAHIR DAR UNIVERSITY

BAHIR DAR INSTITUTE OF TECHNOLOGY

SCHOOL OF GRADUATE STUDIES

FACULTY OF COMPUTING

Approval of thesis for defense result

I hereby confirm that the changes required by the examiners have been carried out and incorporated in the final thesis. Name of Student: Anegaw Sisay Signature ______ Date November 2023. As members of the board of examiners, we examined this thesis entitled "Deep Learning-Based Business Income Tax Fraud Detection Model." We hereby certify that the thesis is accepted for fulfilling the requirements for the award of the degree of Masters of Science in "**Information Technology**".

BoardofExaminers

Name of Advisor	Signature	Date
Dr.Mekonnen Wagaw (PhD)		27/02/2016 E.C
Name of External examiner	Signature	Date
Dr.Getachew Mamo (PhD)		July03,2023
Name of Internal Examiner	Signature	Date
Dr.Abdulkerim Mohammed (PhD)		29/02/2016 E.C
Name of Chairperson	Signature	Date
Mr.Dagnachew Melesew (PhD)		27/02/2016 E.C
Name of Chair Holder	Signature	Date
Dr.Abdulkerim Mohammed (PhD)		29/02/2016 E.C
Name of Faculty Dean	Signature	Date
Asegabeg T.		29/02/16.

Bahir Dar University Bahir Dar Institute of Technology Faculty of Computing

ACKNOWLEDGEMENT

First and foremost, I praise the Almighty God,who favours me to begin and to bringtoanendthis study.Mynextandmostearnestgratitudemustgotomyadvisor Dr.MekonnenWagawfor hiscontinuoussupportandguidancethroughoutthevariousstages of this thesis. Without his personal involvement and intervention at criticalstages, it would have been very challenging to complete the thesis according to theinitially set plan. He provided critical and useful feedback and suggestions onapproachingtheresearchproblemsystematicallyandtactfully.Heshowedmehowthe research will be produced, and how the best advisor interacts with his student(s),and he is really a model for academic professionals. I thank him all the time. I wouldlike to thank my classmates for their all-rounded help during our stay at Bahir DarUniversity. I am also very grateful to my family for their all-rounded support. Thedeepest gratitude also goes to members of the Information Technology ManagementDirectorate Department in MOR. They are offering the necessary data and to theirunreserved professional support and advice. Special thanks also go to all my friendsinsideme.Finally,Iwouldliketothankmywife FentaneshMelkuforherpricelesss upport; yoursupporthelpsmetofinalizethestudy.

TableofContents

LISTOFFIGURES

LIST OFTABLES

LISTOFABBREVIATIONS

ADAGRAD: Adaptive

Gradient**ADAM**:AdaptiveMomentEstima

tion**AI:**ArtificialIntelligence

AUC: Area under the

Curve**ANN:**ArtificialNeuralNetwo

rk**CIO**:ChiefInformationOfficer

CNN: ConvolutionalNeuralNetwork**CS**

V: Comma Separated

Values**DSS:**DecisionSupportSystem

DL:DeepLearning

FNR: FalseNegativeRate

ITMD: InformationTechnologyManagementDirectorate

JSON:JavaScriptObjectNetworking

MLP: Multilayer

Perceptron**ML:** Machine

Learning**MOR:**Ministry of

Revenues**RELU**:RectifiedLin

earUnit

RNN:RecurrentNeuralNetwork

ROC:ReceiverOperatingCharacteristic

SGD:StochasticGradientDescent

SIGTAS: StandardIntegratedGovernmentTaxAdministrationSystem

TIN: TaxIdentificationNumber

TNR:TrueNegativerate**T**

PR:True Positive

rate**VAT:**ValueAddedTa

x

ABSTRACT

The collection of tax is the mainsource of incomefor the government. Taxcollecting has been associated with a lot of fraud, which is a challenge to detect.Fraud involves one or more persons who intentionally act secretly to deprive thegovernment of income and use it for their benefit. This study was initiated to explorethedeeplearningtechnologyfordevelopingmodelsthatcandetecttaxfraudusing dataobtained fromtheMinistryofRevenuesinEthiopia.

To collect the data, the researcher used interviews and observation as primary dataand database analysis as secondary data. The dataset used in this study had beentakenfromEthiopia's Ministryof Revenues.Afterselectingthedataset,pre-processing techniques such as filling missing records, removing outliers, reducingthe dimension, selecting the most relevant features, and finally normalizing thedataset input using features scaling are performed. The deep learning models for taxfrauddetectionareimplementedusingPythonprogramminglanguage.Theexperiments had beenconducted by using the 23536-dataset records.We used **80%**of the dataset for training the model and the remaining **20%** of the dataset for testingthe performance of the model that is developedby the ConvolutionalNeuralnetwork. The model had shown the highest classification accuracy of 84.64%. Thenthis model was tested by 4708 testing datasets and scored a prediction accuracy of84.41%. The results of this study have shown that deep learning technology isvaluablefortaxfrauddetection.

Keywords:Tax,Taxfraud,deeplearning,Keras,CNN

CHAPTER ONEINTRODUCTION

1.1 Background

Taxation is one of the important elements in managing national income, especially indeveloped countries [1]. Taxation is a taxing authority, usually a government, levies, orimposes a tax. The purpose of taxation is “for the maintenance of the public force andadministrative expenses” [2]. The term tax applies to all types of involuntary levies,fromincometocapitalgainsto estatetaxes.

Common classifications of taxes are direct and indirect taxes [3]. A direct tax is a formaland economic incidence that is essentially the same. The taxpayer is not able to passthe burden to someone else. On the other hand, an indirect tax is a tax whereby thetaxpayer’s burden to pay the tax can easily be passed on to another person. Generally,thetaxincidenceofanindirecttaxison theend-user.

In the world, most countries have a taxation system history. From these, Ethiopia hada taxation system history, which began in the 1940s. The modern income tax system ofEthiopia began in 1944 E.C [4]. Ethiopia issued the first income tax law at a time whenEthiopia had a special political relationship with Great Britain, and the EthiopianincometaxschedulestructurewasborrowedfromtheBritishtraditionofincometax [1]schedules.

Ethiopia has issued largely autonomous income tax laws for Petroleum income tax,mining income tax, Agricultural income tax, and Main income tax [4]. The “main”income tax system consists of four schedules, identified by alphabets: A, B, C, and D.Schedule “A” income tax system charges “income from employment”; Schedule “B”“income from the rental of buildings”; Schedule “C” “income from business” andSchedule “D” “miscellaneous income” [4]. In Ethiopia, only the agricultural income offarmers and cooperatives is decentralized to the Regional Governments. Both theFederal Government and the Regional States have issued their income tax laws inrespect ofincomesourcesreservedtoeachrespectivelybytheEthiopianConstitution,although almost all of them are modelled upon the Federal Income Tax Law issued in2002[5].

The tax authorities in Ethiopia categorized taxpayers into three Categories. Category“A” taxpayer’s annual income is more than 1,000,000 Birr. Category “B” taxpayer’sannual income is between 500,000 Birr and 1,000,000 Birr. Category “C” taxpayers’annualincomeislessthan500,000Birr[4].

The Collection of taxes is the main source of income for the government. However,during tax collection, the main problem is getting the exact income report from thetaxpayers. This problem has been directed to the government's annual budget. Theannual expenditureofthegovernmentdependsonincome[6].

Despite technological advancements providing efficiency in conducting business, theseimprovements have also brought about an explosion of data, creating a challenge todetectfraudintax data.

Thelegaldefinitionoffraudvariesfromcountrytocountry.Fraudessentiallyinvolvesusing deception dishonestly to make a personal gain [7]. Whilst the Oxford EnglishDictionarydefines,fraudaswrongfuldeceptionintendedtoresultinpersonalgain[8].In the academic literature, fraud has been defined as leading to the abuse of a profitorganization's system without necessarily leading to direct legal consequences [8].Fraud involves one or more persons who intentionally act secretly to deprive thegovernment of income and use it for their benefit. Fraud is as old as humanity itself andcan take an unlimitedvarietyof differentforms [9]. Parties andorganizations to securea business advantage through the unlawful act to obtain money, property, services, orto avoid payment perpetrate fraud. Traditional ways of data analysis have been in useforalongtimeasamethodofdetectingfraud.Theyrequiretime-consuminginvestigations that deal with different domains of knowledge like finance, economics,business practices, and law [9]. Taxpayers to reduce tax liability mostly perform taxfraud, this illegal action performed to misrepresent the financial facts to governmentandtaxauthoritiesbyprovidingfalsetaxreporting[10].

This study focuses on the historical data of business income tax fraud caused bybusiness taxpayers in a particular tax fraud report that may be declaring less income,lessprofit,exaggeratedcosts,misrepresentationoftheprice,andtocomplexnetworksoffinancialtransactions(openmorebranches).

This study detects business income tax fraud by using Deep Learning technology. DeepLearningisaclassofmachinelearningalgorithmsintheformofaneuralnetworkthat uses a cascade of layers (tiers) of processing units to extract features from data andmakepredictionsaboutnewdata[11],[12].

Deep learning is used for many applications like fraud detection on tax data, plagiarism,computer network management, and event detection to name a few [11], [12].

ThereareavarietyofDeeplearningnetworkssuchasMultilayerPerceptron(MLP),AutoEncoders (AE), Convolution Neural Networks (CNN), and Recurrent NeuralNetworks (RNN). Besides, Deep Learning supports different libraries and frameworks[13], [14] such as Keras, TensorFlow, Pandas, Sklearn and Numpy, MatPlot, andSeaborn,etc.

This study used Pandas libraries to read Excel files, and MatPlot libraries to plotdifferent graphs and the study used Convolution Neural Network (CNN), which,significantly, enhances the capabilities of the feed-forward network such as MLP byinserting convolution layers. In addition, thestudy used the Keras framework toimplementtheCNNnetworksusingthePythonProgrammingLanguageintheAnacondaenvironment.

1.2 Motivation

A principal motivation behind this study, the global economic crime survey of 2016suggests that more than one in three (36%) of organizations experienced a taxdeception problem. The taxpayers do not pay tax properly that is in a year, morethan seventy-eight (78 %) percent, and the numbers of taxpayers who commit taxfraud become increasing year-to-year accordingly MOR reports in Ethiopia.

Thesedishonesttaxpayers’activitieshaveanegativeinfluenceonhonesttaxpayers.Also,below twenty (20%) taxpayers should be audited yearly from the total taxpayers,which is prone to fraud based on auditor analysis. All taxpayers should be auditedwithinfiveyears, but according to theauditors' report, theycannot audit alltaxpayers. One of the measures of a country’s tax system is GDP at the currentmarket price or tax/GDP ratio. The performance of the Ethiopian tax system has notimproved quite considerably over the last decade. In 2015 shows that, the tax-to-GDPratiohasbeenfarlowerthaneventheSub-Saharanaverage.Ethiopiancurrenttax-to-GDP ratio of 11% is far lower than the average for developed tax systems(25-35%),developingcountries(18-25%),andeventheSub-Saharanaverage(16%)

[1].

ThegovernmentofEthiopiatriestominimizethefraudsters'taxpayersbyaddressingdifferent techniquessuch asgiving ashort training awarenessfortaxpayers and traditional auditing techniques. However, there is still a challenge inthe tax system. The above-mentioned illegal activities in this Section motivate us tostudythesefraudstertaxpayersbyusingaDeepLearningalgorithmtofacilitatethemanagementandaudittasksofthe organization.

1.3 ConceptualFramework

In this study, the conceptual framework comprises the basic components of the studyas well as the relationship of these elements with one another, which is used as aspringboard by the study that explains the stages or steps done in the process. Thecomponentsaredatasource,dataprocessingtool,modeldevelopmenttechniqueoralgorithms,andmodelevaluationtechniquesasdescribedinFigure1.1.

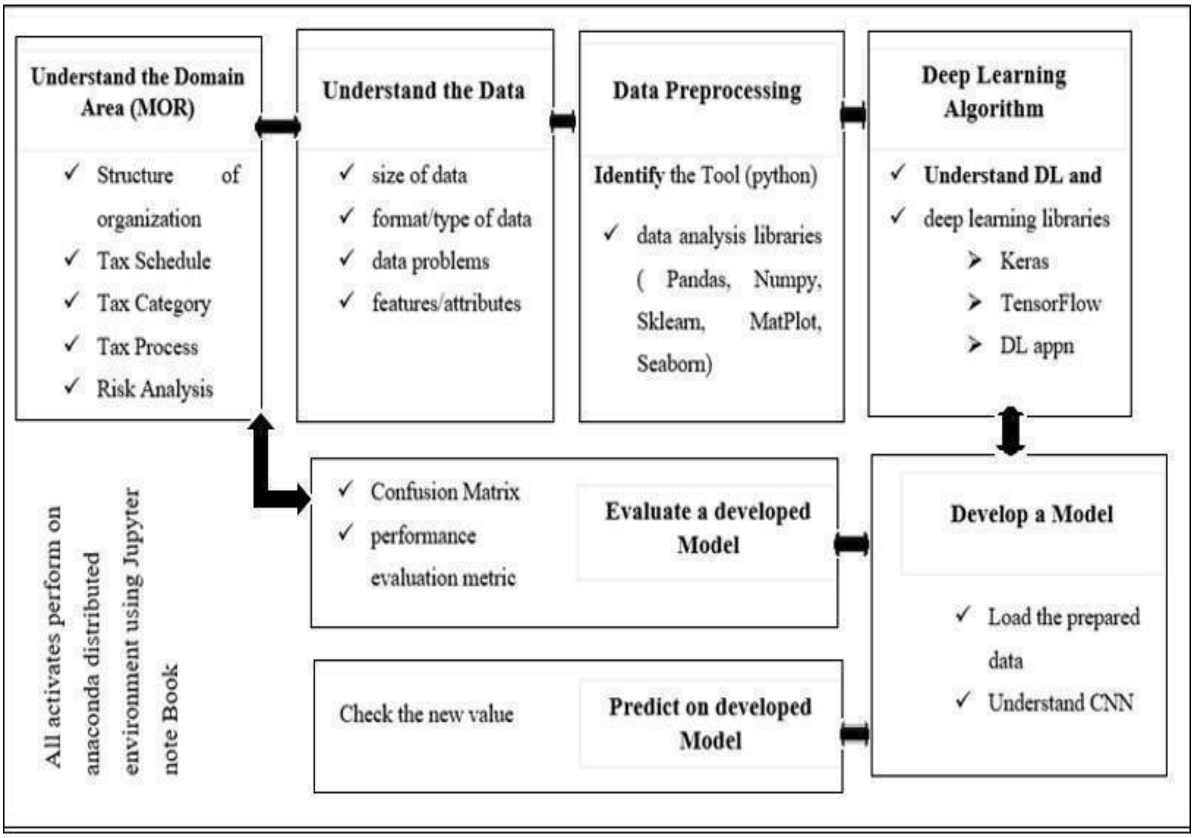


Figure1.1TheoreticalframeworkoftheStudy

1.4 Statementoftheproblem

Tax fraudis an intentional reduction ofthetax liabilitystemmingfrom realtransactions [7]. Tax fraud typically includes underreporting profits (Gross profit)andAnnualincomesales,overstatingdeductions(expenditure),underreporting employee wages, failure to register tax statements, hiding of taxable receipts comingfrom the production and distribution of real products and services (withholding),overvaluingofVATspenton inputs and abuseoftax return throughuntruetransactions [16]. These problemschallenging the governments to collect tax,especially in developing countries, have been associated with a lot of fraud, whichis a challenge to detect. In Ethiopia, the tax administration is not an exception tosuchchallenges.

Despite putting up various audit techniques and strategies to fight tax fraud, such asdesk audits, spot audits, comprehensive audits, and special audits. Tax Fraud hasbeen continuing to be a challenge because fraud remains a limiting factor to thecapacity of the government in raising revenues to carry out economic policies.Traditional strategies of auditing, which are investigating audits and tax audits usingriskanalysiscriteriacannotfixthelossamountoftaxableincomeofthegovernment. Recentresearcherstendtousesimilarandstandardmethodstodetecttaxfraudandinformation on taxes is being stored in messy formats. Deep learning is needed toaddressincome frauddetectionbecauseitcanlearntoidentifypatternsindatathatwouldbedifficultorimpossibletoidentifyusingtraditionalmethods.Forexample,deep learning models can be trained to identify patterns in tax returns that areassociated with fraud, such as unusual spikes in income or expenses, or the use ofcomplex financial arrangements. Here are some of the advantages of using deeplearningforincomefrauddetection:

Accuracy: Deep learning models can be very accurate in detecting fraud. In fact,deep learning models have been shown to be more accurate than traditional methods,suchasrule-basedsystems.

Scalability: Deep learning models can be scaled to handle large amounts of data.This is important because income fraud is a growing problem, and the amount ofdatathatneedsto beanalysedisincreasing.

Cost-effectiveness:Deep learningmodels can becost-effective.The cost of traininga deep learning model can be high, but the cost of preventing fraud can be muchhigher.

However, there are also some challenges associated with using deep learning forincomefraud detection:

Data requirements: Deep learning models require large amounts of data to train.Thisdatacanbedifficultandexpensivetocollect.

Complexity: Deep learning models can be complex to build and maintain. Thisrequiresspecializedskillsandresources.

Interpretability:Deeplearningmodelscanbedifficulttointerpret.Thiscanmakeitdifficult to understand why a model has flagged aparticular piece of data assuspicious.

1.5 ObjectivesoftheStudy

1.5.1 GeneralObjective

The general objective of this study is to design a deep learning-based businessincometax frauddetectionmodel.

1.5.2 SpecificObjectives

Thespecificobjectivesofthisstudyare:

- Toanalysethefraudulenttaxpayersonbusinessincometax.
- Toselectanappropriatemethodologyandtoolstoconstructatargetdataset.
- TodesignanddevelopproposedModel.
- TomeasuretheperformanceoftheproposedModel.

1.6 ResearchQuestions

Thisstudyansweredthefollowingresearchquestions.

- Whataretheimportantparametersthatinfluencetheidentificationoffraudulent taxpayersonbusinessincometaxes?
- Howcandeeplearningmodelsbeusedtodetectnewandemergingformsofbusinessincometaxfraud?
- HowDeeplearningtechniquescanbeappliedtodetectfraudonbusinessincometaxtoimprovethequalityofserviceandminimizefraud?

1.7 MethodologyoftheStudy

To define the research problem properly, primary data collected by interviewingconcernedexpertsaswellasthroughobservation(questionsroseduringthe interview described in Appendix G). Relevant literature reviewed on MachineLearning,DeepLearningalgorithms,andfrauddetectionontaxdata.Thestudyused mixeddatacollectionmethodsandtechniquestosplitthedependentandindependentvariables,whichhaveanequalchanceforthepopulationtoselect.

In this study, python is empowered to implement most of the technical aspects ofthe data pre-processing tools within a deep learning algorithm to develop a model.In this study, the undertaken activities are data collection, data pre-processing,model building, and model evaluation and prediction. To implement the proposedstudy, software tools such as TensorFlow, Jupyter Notebook, Pandas, Numpy,Sklearn,Matplotlib,Seaborn,andKerashavebeenusedandAnacondaenvironmentshavebeenused.

1.8 ScopeandLimitationsoftheStudy

1.8.1 Scopeofthestudy

The scope of this study focuses on business income tax (Schedule C) taxpayers whoprepared financialstatements andbalance sheetsforfederalgovernments.Thefinancial statements and balance sheet are analysed by using tax risk analysingactivities or criteria such as loss declaration, late payment, profit margin, custom,commencement, asset, auditoption, and intelligence to name afew. Tax riskanalysisis the most effective working area of the Ministry of Revenues. The effectiveness oftax risk analysis is used to improve tax revenue performance, to identify auditmethods, and it used to detect taxpayers from fraud. All taxpayers' files should beregistered based on the criteria of tax risk analysis. Based on the scope we had donedifferent activities for this study. We had identified the criteria of tax risk analysisthat were used to minimize tax fraud. We had analysed the taxpayers' informationbased on these identified criteria to prepare the target dataset for MOR on businessincometax. Also,wehadanalysedthedatapreparationmethods,wehadidentifiedadeeplearning-basedalgorithmtodesignafrauddetectionmodelonbusinessincometax,and wehadconstructedamodeland wehadimplemented differentexperimentsusingCNNwithKeraslibrary,whichhelpstoidentifytaxfraudsters.

1.8.2 LimitationsoftheStudy

This study is limited to Schedule "C" taxpayers who are paying taxes to the FederalGovernment and the study only uses convolutional neural network algorithms. Thestudy is also limited on criteria of tax risk analysis based on the nature of the study,time,andresourceconstraintsliketheabsenceofappropriatedata.

1.9 SignificanceoftheStudyandBeneficiaryofStudy

This study facilitates management and audit activities of the tax process in theMinistryofRevenuesbecausebothmanagementandaudithaverolestoplayinthedetection of fraud. After analysis, the risk analysis processes of both managementand audit can detect tax fraudster taxpayers, which have an expectation to save time,increaseincomeforthegovernment,andusedtoachieveadevelopmentplan.

This study has paramount uses for different stakeholders who are interested in thetaxation system. The outcome of this study was used as a benchmark for auditors aswell as a source of a methodological approach for dealing with deep learning onfraudmanagementaswellasothersimilarareas.

Finally, the study might have invaluable importance forfuture researchers who needto conduct a study. Taxes are fundamental to the existence and give the governmentpower to allocate resources; to enable the government to provide/support socialdevelopment; to stabilize the economy; to constitute and define the marketplace;andtoencourageoptimaleconomicgrowth.An improvedtaxsystemimprovestherevenues available for supporting public service without increasing the current taxburden on compliant taxpayers. Moreover, an improved tax system bolsters citizens'satisfaction by increasing their faith in the system and promoting the perception thateveryonepaystheirlegalshare.Understandingtheproblemfacingthetaxadministration system is the major factor that contributes to the success of theoverall tax system. Unless the problems are pointed out and addressed properly, itmaybedifficulttodesignasufficientandeffectivetaxsystemthathelpstonarrowtheexistingtaxadministrationgap.

1.10 EthicalConcern

The data for this study is obtained from the Ethiopian Federal Tax Authorised officebypresentingacooperationletterfromBahirDarUniversity.Theletterwasdirectedto the Chief Information Officer (CIO) of the department and I got the data from theMinistryofRevenue.

1.11 TheStudyOutlines

This study is structured into six chapters. Chapter 1 deals with the introduction ofthe whole document. It states the statement of the problem, the objective of thestudy,research methodologyused.

Chapter 2 describes the state of the literature review and related works and the thirdchapterpresentsthemethodsandproceduresofthestudyusedanddatasetpreparation, software tools, and performance evaluation metrics. The fourth Chapterpresents the Design and Implementation of the data. Chapter 5 presents the modelExperimentationandDiscussiononresults.Thelastchapterpresentstheconclusions andRecommendations.

CHAPTER TWOLITERATURERE VIEW

2.1 Introduction

Thischaptermainlyfocusesonthebackgroundinformationandreviewofliteratureof the domain of this study. It includes a detailed explanation of taxation systems,tax fraud detection approaches, machine learning, and deep learning algorithms, andrelated works. Finally, the chapter is concluded with a summary of related worksandthemaingapsthatshould besolvedinthisstudy.

2.2 TaxationSystem

A tax is not a voluntary pay mentor donation, but a required, according to legislativeauthority [17]. Tax collection is performed by a government revenue agency suchas Canada Revenue Agency, the Internal Revenue Service in the United States,Kenya Revenue Authority, and Ghana Revenue Authority [4]. Tax involves everyaspect of income-generating activities and consumption items, and requires not onlythe administrative capacity of revenue authority but also the involvement of privatesectors through proper accounting and reporting [18]. The classification of tax iscategorized into two [19]. These are direct and indirect taxes as defined in Chapter1 in Section 1.1. The description of each tax category explains as follows in table2.1

Table2.1DirectandIndirectTaxTypes

IndirectTaxes	**Description**
ValueAddTax(VAT)	Toaproductfromabusinessisthesalepricechargedtoitscustomer,minusthecostofmaterialsandothertaxableinputs.
TurnoverTax	Itisanindirecttax,typicallyonanadvalorembasis,applicabletoaproductionprocessorstage.
ExciseTax	It is an inland tax on the sale, or production for sale, of specific goods; or,more narrowly, as a tax on a good produced for sale, or sold, within acountryorlicensedforspecificactivities.

DirectTaxes	Description
WithholdingTax	Isagovernmentrequirementforthepayerofanitemofincometowithholdor deducttaxfromthepayment,andpaythattaxtothegovernment?
PersonalIncome Tax	Everypersonderivesincomefromemploymentorotherprivateorganizationsor non-governmentorganizations.
RentalTax	Ataxthatisimposedontheincomefromtherentalofbuildings.
CostSharing	Aportionofthetotalprojectorprogramcostsrelatedtoasponsoredagreementth atiscontributedbysomeoneother thanthesponsor.
BusinessProfit Tax	A tax is imposed oncommercial, professional,orvocational activityoranyotheractivityrecognizedastradebythecommercialcodeoftax .
ScheduleD-GamesOfChance	Everypersonderivingincomefromwinningatgamesofchance/forexample,lot teries, tombola,andothersimilaractivities.

2.2.1 TaxationSysteminEthiopia

An income tax is one of the main sources of Federal and Regional Governmentrevenues. The Ethiopian government used income taxes as one of the principalsources of domestic government revenue since the beginningof modern taxation [4]in the1940s.

The Ethiopian income tax system is a “scheduler” in structure and orientation, thecomputation, assessment, and collection of income taxes based on some identifiedsources of income, like income from employment, income from the rental ofproperty, and income from business. The modern income tax system of Ethiopiabegan [4] in 1944 E.C. when the first income tax law was issued to levy a tax on theincome of individuals and businesses. The first income tax law was scheduled ashaving successive income tax laws issued over the years. Ethiopia issued its firstincome tax law at a time when it had special political relationships with GreatBritain, and its scheduler income tax structure was borrowed from the Britishtraditionoftaxingincomeby schedulesorsources.

The contents of the “schedules” of Ethiopian income tax have changed throughsuccessive income tax reforms in Ethiopia. Some of the original schedules haveeithercompletelydisappearedor been replaced byothers,while some of thescheduleshaveretainedtheiroriginalcontents[21].

The old income tax proclamation 286/2002 is amended to the federal income taxproclamation 979/2016[6]. The proclamation provides for the taxation of income inaccordance with the schedules: Schedule ‘A’ income from employment, ScheduleB income from the rental of the building, Schedule C income from a business,schedule D other income, and exempt income (Federal income tax Proclamation No,979/2016)[6].

Income tax shall mean every sort ofeconomic benefit includingnon-recurring gainsin cash or in kind from whatever source derived and in whatever from paid, credited,orreceived[20].

2.3 Taxfraud

Tax fraudis anintentional reductionofthe tax liabilitystemmingfrom realtransactions[21].However,inmanycountries(especiallydevelopingandtransition alcountries), audit performance is reported as aweak aspect of taxadministration, other irrespective aspects are working well [11]. Several developingcountries do not yet have effective audit programs due to insufficient numbers ofthe required highly skilled and appropriately paid audit practitioners, absence of asoundinstitutionalauditpractice,illegalcooperationbetweentaxpayersandauditors, lackof clear politicalsupport for the tax administration, and the deficiencyof an appropriate legal and judicial environment [10]. Additionally, these countriestend to offset weak tax audits by adopting complex procedures, such as increasedfiling requirements and massive cross-checking. The audit is not a very welcomeprocedure forboth the taxpayers and the economy. Conducting audits involves coststo the tax department as well as to the taxpayer. Tax administration agencies shouldusetheirscantresourcesveryjudiciouslytoachievemaximaltaxpayercomplianc e,and minimal intrusion and costs. Among others, having an effective tax auditprogram is a key success factor for cost minimization and detection of tax fraud aswellasproactively preventingtaxfraud [7].

2.4 TaxFraudDetectionApproaches

2.4.1 MachineLearningApproaches

Machine learning is an application of AI that makes a machine learn and improveautomatically without being explicitly programmed [22]. Unlike classical computerprograms that perform a task explicitly programmed by the programmer, a machine-learning program uses a generic algorithm that can give information about a set ofdatawithouthavingtowriteanycustom program,whichisspecifictotheproblem.That is instead of writing anew program for thespecific problem, we only feeddatatothegenericalgorithmanditcomputesthatdata thenthealgorithmbuildsitsownlogic based on the given data [23]. The goal is to allow the computer to learnautomaticallywithoutthehelpofhumanbeingsandadjustaccordingly.
In this study for the tax fraud problem, the training dataset labelled as fraud and NonFraud were used. After learning from the dataset, the algorithm is able to predictwithanunseendatasetduringthetraining.

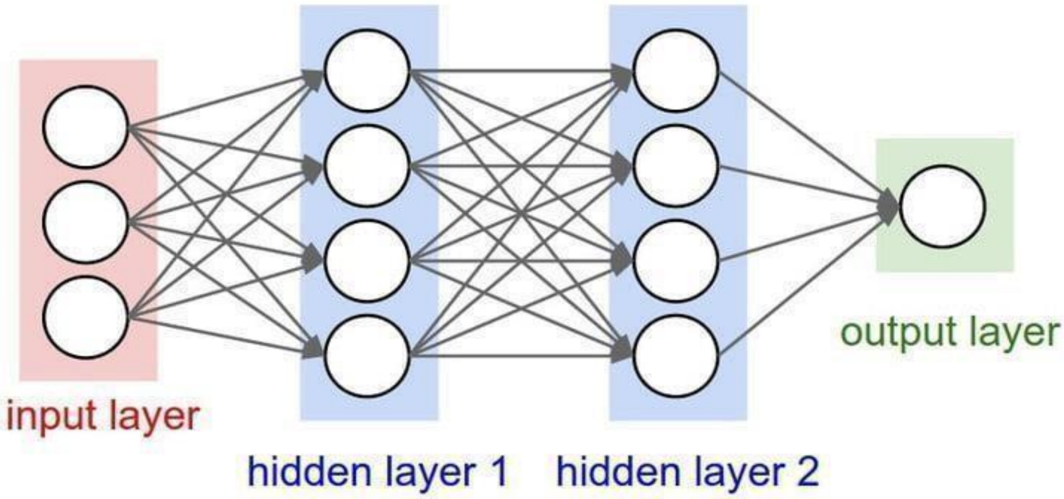
The second main category of machine learning is unsupervised (descriptive) learning,this approach has little or zero knowledge of the output and we want to try to findpatterns or groupings within the data. The goal is to find an interesting pattern or tomodeltheunderlyingstructureinthedatainordertolearn moreaboutthedata[23].

2.4.2 ArtificialNeuralNetwork

Artificial Neural Network(ANN) is one of themostwidely usedsupervisedmachine learning models. The primary focus of this study is a special type of NN.ANN sometimes called neural networks, computer program developed to mimic thehumanbrain[13].Theterm“neuralnetwork"originatedin1943tofindamathematical representationofbiologicalinformationprocessing[27].Likehumans, ANNs are trained through experience by giving appropriate exampleswithout any special programming. ANNs are excellent at finding patterns that areverycomplexforhumanstoextract.Theygainknowledgebycollectingrelationships and patterns in the data that is provided during the training [21, 23].ANN contains multiple layers, where each layer will have a number of neurons. Aneuron is a smaller building block of the network and it accepts an input, appliessomecomputation,andgeneratesauniqueoutput[13].

2.4.2.1 Multi-LayerNetworks

ANNs are a combination of multiple artificial neurons grouped in layers [13, 21].Most of the ANNs except single-layer networks (a network without a hidden layer)have three types of layers, the input layer, one or more hidden layers, and the outputlayer. Multi-layer networks have one or more hidden layers. Each of the layers inthe network consists of one or more neurons. The neurons in the input layer acceptinformation from outside the network and transfer it to the hidden layers of thenetwork. The input layer passes the data without modification (no computation isperformed) process. The hidden layers (sometimes called layers with neither outputnorinput)performmathematicalcomputationandtransfertheinformationfromth einput layer to the other layer. Most of the computation in the network is performedin the hidden layer. Neurons in the output layer perform computation and transfertheinformationtooutsidethenetwork.Theoutputlayer transfersactivationsinthehidden layer to actual output, for example, classification and prediction. Multi-layernetworks(ormulti-layerperceptions)arealsoknownasfeed-



forwardneuralnetworks.

Figure2.1ExampleofMultilayerNetwork

As shownInFigure2.1[29]above,eachoutputofalayer oftheneuronisreceivedasaninputineachlayerofthenextlayeroftheneuron;thiskindofneur alnetworkis called a fully connected feed-forward neural network. In this type of neuralnetwork, neurons in the input layer receive the original input data while otherneurons in the other layer receive the outputs of previous neurons. In a feed-forwardneuralnetwork,informationflowsfromtheinputlayerto theoutputlayerthrough

the hidden layer without going back. Each neuron in the network has an equalnumberofweightstothenumberofneuronsinthepreviouslayer[27].

2.4.2.2 BackpropagationAlgorithm

The backpropagation algorithm allows theinformation to flow in reversedirection,the information flows backward from the output neurons to the input through thehidden layers in order to compute the gradient [24, 20]. During the training of theneural network, weights are selected appropriately; therefore, the network learns topredict the target output from known inputs [30]. Even thoughcomputing theanalyticalexpressionfortheweightsoftheneuronsisstraightforward,itiscomputationally expensive. Therefore, we need to find a simple and effective deeplearning-based fraud detection model for the tax system in Ethiopia algorithm,which helps us to find the weights. The backpropagation algorithm provides asimpleandeffectivewayforsolvingtheweightsiterativelyinordertoreduceerror(minimizingthedifferencebetweentheactualoutputandthedesiredoutput)intheneural network model [22,30]. Small random values have been initialised for theweights of the network neuron when an input vector is propagated forward to theneural network. By using a loss function, the predicted output (output of thenetwork)andthedesiredoutputs(outputfromthetrainingexample)arecompared. i.e. the gradient (error value of the network). The error value is simply the differencebetweentheactualoutputandthedesiredoutput.Theerrorvaluesarethenpropagated back from the output layer to the input layer through the hidden layersand then the error values of the hidden layers are calculated. In this process, theweightsofthehiddenlayersareupdated.Thisiscalledlearningduringthetrainingprocessoftheneuralnetwork.Whentheweightsareiterativelyupdated,theneuralnetwork gets better. The algorithms continue this process by accepting new inputsuntiltheerror valueislessthanthelimitvalueoftheweightwesetbefore[20].

2.4.2.3 ActivationFunction

The final output of each neuron in the neural network is determined by activationfunction φ. Activation functions are functions that decide whether a neuron shouldactivate (fire) or not by calculating a weighted sum and adding bias with it [31].Activation functions introduced non-linear properties to the NN to overcome thedrawback of early neural networks (Perceptron). The drawback of early NN was theproblemofcomputingnonlinearandcomplexproblems.Themainpurposeofthe activation function is to convert the input in a neuron of NN to output. The outputof thatneuron is used as an input in another neuron of the next layerof the network.Ifwedonotuseactivation,theoutputoftheneuralnetworkwillbesimplyalinearf unction. A linear function is not applied in algorithms that need to learn fromcomplex functional mapping on data [32]. The main reason that makes us use non-linearity is that we want the NN model, which learns and represents any arbitraryfunction,whichmapsinputsfromtheoutput.

In this study, the most widely used activation function, which is called RectifiedLinear Unit (ReLU), has been used in the hidden layer of the network to make ourmodelmorepowerfulandtolearncomplexfeaturesfromdata.Itisusedtocreatealight weight and effective nonlinear network [22, 33]. ReLU became popular in thepast few years and now it is a state-of-the-art activation function for hidden layers[24,20].

The main reason that makes ReLU simple and efficient is that it activates some oftheneuronsat atime.i.e.iftheinputisnegative(x<0),itconvertsittozeroandtheneuron is not activated. ReLU can’t be applied in the output layer of the neuralnetwork and this is the main drawback of this activation function. The sigmoidactivation function has been used for the output layer of the model. The sigmoidactivation function is the best activation function for binary classification and itexists between 0 and 1 [20]. It is the best choice for models that have probabilityoutput since the probability of anything exists between 0 and 1. Unlike the SoftMaxactivation function, the sum of the output of sigmoid functions is not equal to 1.TheSoftMax function accepts arbitrarily n inputs and it gives n output values within arange between 0 and 1. This shows the probability of different classes defining eachinput. The sum of the value of the output is always equal to 1. SoftMax is the bestchoice of activation function for neural network models that are built for multiclassclassification[11].

2.4.3 Deeplearning

Deep learning is a subfield of machine learning that uses a neural network for itsarchitecture and its learning is based on a data representation algorithm instead oftask-specificalgorithms[34,24,].Inthelastdecade,neuralnetworkapplicationsisgrowing faster than ever mainly because of many powerful computers (inexpensiveprocessingunits suchasGPU) andalargeamountofdata.AsdiscussedinSection

2.4 above anANNhasoneormoreprocessinglayers.Dependingontheproblemwe want tosolve, the numberof layerswe use in the networkdiffers. If the numbersof layers are two or three we call the network shallow architecture. When an ANNarchitecture that contains a very large number of layers, the network is called deeparchitecture and deep learning refers to this deep architecture of NN [35].Multilayernetworkswereknownsincethe1980s,butforseveralreasons,thenetworks were not used to train a neural network with multiple hidden layers [22].The main problem thatprevented the use of multilayernetworks at thattime was thecurseofdimensionality,i.e.ifthenumberoffeaturesofdimensiongrows,thenumberofconfigurationsincreases.
Asthenumberofconfigurationsincreases,thenumberofdatasamplesforthetrainingincreasesexponentially.Therefore,collectingsufficienttrainingdatasetswastime-consuminganditwasnotcost-effectivefortheusageofstoragespace[22,36].Nowadaysmostoftheneuralnetworks areoftencalleddeepneuralnetworksandtheyarewidelyused.Wecantrainaneuralnetworkwithmanyhiddenlayers becauseahugeamountofdata,aswell asstoragespace,andcomputational resources,isavailable.
Thetraditionalmachine-learning algorithm needs separatehand-tuned feature extraction before the machine-learning phase. Deep learning has only one neuralnetwork phase. At the beginning of the neural network, the layers are learning torecognize the basic features of the data, and that data feeds forward to the otherlayersinthenetworkforadditional computationofthenetwork[22].
Deep learning techniques are new and rapidly evolving. Nowadays deep learningperforms better than other traditional machine learning approaches because of theavailability of a large amount of data and high-performance computing machinecomponentssuch asGPU[24].

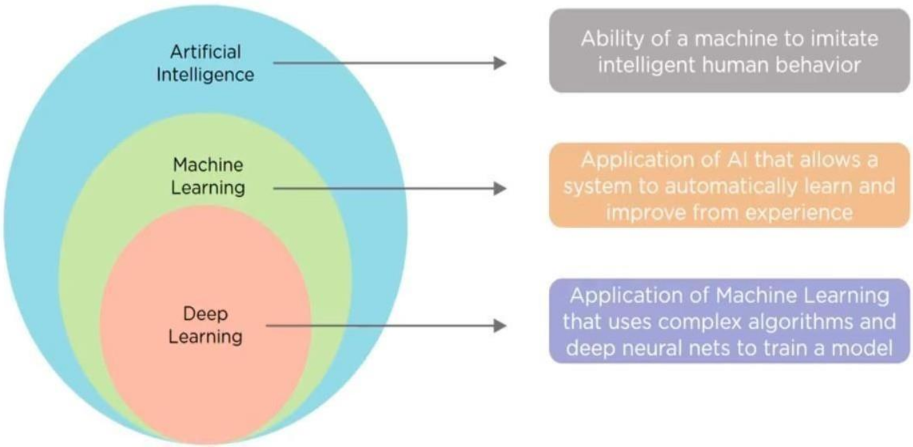


Figure2.2DiagrammaticrelationshipsofAI,MLandDL

Deep learning methods use multilayer processing with better accuracy performanceandunliketraditionalmachinelearningapproachthereisnoexplicitfeatureextraction, i.e. in deep learning architecture features are extracted automatically fromthe raw data and we can perform feature extraction and classification (it might berecognition depending on our problem) at once, therefore we only design a singlemodel.

To overcomethecomplexityofthedesign,deep learningmethods usebackpropagation algorithms, loss functions, and too many parameters that make themodeltolearncomplexfeatures.Theparametersare:

Dropout

Dropout is a weight regularisation in neural networks to avoid overfitting the data.Typically, the Dropout is 0.8 (80 % of neurons present randomly all the time) in theinitiallayersand0.5inthemiddlelayers[27].

OptimizerandLearningRate

Optimizer is used to optimize learning rates by using various techniques [28]including:

- Stochastic Gradient Descent (SGD): Gradient descent is a way to minimisean objective function parameterized by a model’s parameter by updating theparameters in the opposite direction of the gradient of the objective function.Stochastic Gradient Descent (SGD) and find the best solution. If the networklearns very fast, it may find suboptimal solutions if it learns very slow; it willtakevery longtotrain anetwork [13].

- ❖ Nesterov Accelerated Gradient (NAG): If a ball rolls down a hill and blindlyfollows a slope, it is highly unsatisfactory and it should have a notion ofwhere it is going so that it knows to slow down before the hill slopes upagain.NAGisawaytogivemomentumtothiskindofprescience[29].
- ❖ Adagrad(Adaptivegradient)isanalgorithmforgradient-basedoptimization that adapts the differential learning rate to parameters, performing largerupdatesfor infrequentandsmaller updates forfrequentparameters.
- ❖ Adadelta:Adadeltaisanextensionof Adagradthatseekstoreduceits aggressive,monotonically decreasing learning rate. Instead of accumulatingall past squared gradients, Adadelta restricts the window of accumulated pastgradientstosomefixed size.
- ❖ RMSprop:RMSpropandAdadeltahavebothdevelopedindependently aroundthesametimetoresolveAdagradradicallydiminishinglearningrates
- ❖ Adam(AdaptiveMomentEstimation):Adamisanothermethodthatcomputes adaptive learning rates for each parameter. In practice, Adamgivesthebestresults.

Loss Function: To compute the error between actual and prediction values andmeasure the model's performance. Hyper parameters are fine-tuned to minimize theloss function. Some common loss functions are- Mean Square Error, Log loss, andCrossentropy.

Epochs: One completes a set of feed forward and backpropagation to train the entirenetwork.Onepassesthroughallofthe rowsin thetrainingdataset.

Batch Size: No input observation that is processed in one epoch. One or moresamples are considered by the model within an epoch before weights are updated.One epoch consists of one or more batches, based on the chosen batch size and themodelisfitformanyepochs.

Model building: is a key objective of data analysis applications [27]. In the past,suchapplicationsrequired onlyafewmodelsbuiltbyasingledataanalystasmoredata has been collected, and real-world problems have become more complex, it hasbecome increasingly difficult for that data analyst to build all the required modelsand manage them manually [26]. Building a system to help data analysts constructandmanagelargecollectionsofmodelsisapressingissue.

Supervisedvs.UnsupervisedModels

The models are trained using supervised models and Unsupervised Models methods.Supervisedmodelsaretrainedthroughexamplesof aparticularsetofdata,unsupervised models are only given input data and do not have a set outcome theycanlearnfrom.Supervisedmodelshavetaskssuchasregressionandclassification;unsupervised models have clustering and association rule learning. SupervisedModels have algorithms such as MultilayerPerceptron,ConvolutionalNeuralNetworks, and Recurrent Neural Networks, and Unsupervised Models have Self-OrganizingMaps,BoltzmannMachines,andAutoEncoders[30].

Someofthemostcommonlyuseddeeplearningarchitecturesare Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), CNN, DeepBeliefNetworks(DBN),and AutoEncoders.

- RNNisoneofthefirstdeeplearningarchitecturesthatgivesaroadmapto developotherdeeplearningalgorithms.Itiscommonlyusedinspeechrecognition and natural language processing [38]. RNN is designed to recognizethe sequential characteristics (remember previous entries) of the data. When weanalysed time series data, the network has memory (hidden state) to storepreviously analysed data. To perform the present task RNN needs to look at thepresent information (short term dependency) and this is the main drawback.RNN differs from a neural network in that RNN takes a sequence of data definedovertime[38].
- LSTMisaspecialtypeof RNN,whichisexplicitlydesignedtoovercomethe problem of long-term dependencies by making the model remember values overarbitrarily time intervals. The main problems of RNN are vanishing gradientsand exploding gradients. The gradient is the change of weight with regard to thechange in error. It is well suited to process and predict time series given timelags of unspecified duration. For example, RNN forgets the model if we want topredictasequenceofonethousandintervalsinsteadoften,butLSTMrememberssuchkindsofactivities.The mainreasonthatLSTMcanrememberits input in a long period is that it has a memory that is like memory on acomputer that allows the LSTM to read, write and delete information [39]. It ismostly applied to natural language text compression, handwritten

recognition,speechrecognition,gesturerecognition,andimagecaptioning.

- ❖ CNN is the popular deep learning architecture for different fraud detection tasks,especiallyfortaxsystems.
- ❖ DBNisaclassofdeepneuralnetworkswithmultiplehiddenlayerswhereeachlayerof thenetworkis connectedtoeachotherbuttheneuronsinthelayers arenot connected to each other. The training of DBN occurs in two phases. It iscomposedoflayersofRestrictedBoltzmannMachines(RBMs)fortheunsupervise d pre-training and feed-forward network for the supervised fine-tuning phase. During the training of the first phase (pre-training), it learns a layerof features in the input layer. After the pre-training is completed, the fine-tuningphasebegins.Inthefine-tuningphase, itacceptsthefeaturesoftheinputlayer as input and learns features in the second hidden layer. Then backpropagationor gradient descent is used to train the full network including the final layer [40].DBN is applied in image recognition, information retrieval, natural languageunderstanding,andvideosequencerecognition.
- ❖ AutoEncodersareaspecifictypeoffeed-forwardneuralnetwork,whichis designed for unsupervised learning, i.e. when the data is not labelled. The inputsand outputs of AutoEncoders are the same. It accepts and compresses the inputinto alowerdimensionalcode andthenreconstructstheoutputfromthecompressed code. AutoEncoders have three components namely the encoder,the code, and the decoder. The encoder accepts the input and produces output,whereas the decoder produces output by using the code. Anomaly detection isoneofthemostpopularapplicationsofAutoEncoders.

2.5 ConvolutionalNeuralNetwork

A more capable and advanced variation of classic artificial neural networks, aConvolutional Neural Network(CNN) is builttohandle a greater amount ofcomplexity around pre-processing, and computation of data. In this study, CNN isused to detect business income tax fraud by giving a target dataset of fraud and Nofraud taxpayers’ data as an input. For the process of classification, CNN is usedwhich is composed of various sequential layers and every layer of the algorithmtransforms one volume of activation to another using different functions [29]. Thebasic and commonly used layers of CNN are the convolution layer, the poolinglayers,andthefullyconnectedLayer[22].

A. ConvolutionLayer

The main objective of the convolution layer is to extract useful features from theinputdata.Theconvolutionlayerisformedfrom acombinationofasetofconvolutional filters (feature detectors) which are small matrix values with size like3×3,9×9, and so on [29]. The filters are treated as neuron parameters and arelearnable. Every filter is smaller than the input volume in spatial size (width andheight) and extends the depth equal to the input volume (input data). For example,atypicalfiltermighthaveasizeof5×5×3(5widths,5heights,and3depthsforthet hree-colorchannels).

The convolution operation is performed by sliding the filter on the input data fromleft to right across width and height and computes the dot product between the filterand the input data at any position. The output of this operation is called a featuremap(activationmap).Therefore,thefiltersareusedto extractusefulfeaturesfromthe input data. Whenever the values of the filters are changed, the features that areextracted or the feature map also changes. In the following illustration (Figure

1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	1	0	0

Input

1	0	1
0	1	0
1	0	1

Filter / Kernel

2.4)wehaveprepared2Dinputdataofsize5×5and3×3kernels.

Figure2.3ExampleoftheInputvolumeandfilter

The input and the filter were given; the next step is to perform a convolutionoperation by sliding or convolving the filter over the input. At every location, thedot product (by performing element-wise matrix multiplication and summing theresult)iscomputedandstoredina newmatrixcalleda featuremap(Figure2.5).Aswe can see in thefollowing illustration,the outputof the firstconvolution operationis 4 and the second is 3, these results are added to the feature map. The whole processisperformedbyslidingthefiltertotherightandaddingtheresulttothefeaturemap.

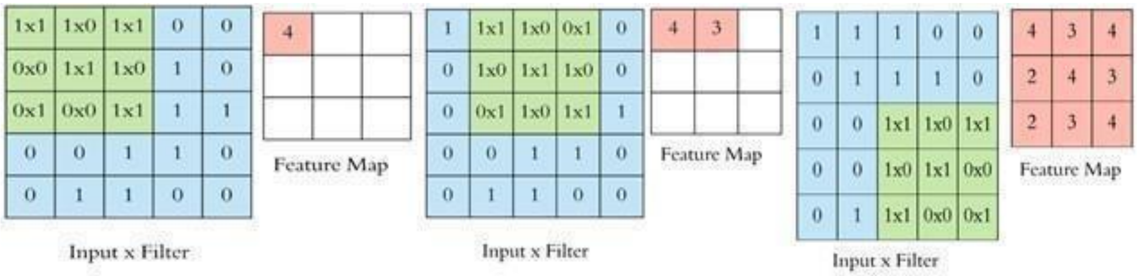


Figure2.4ExampleoftheConvolutionoperation

The area where the convolution operation is performed is called the receptive fieldanditssizeis3×3becauseitisalwaysthesameasthesizeofthefilter.Weperformas many convolution operations as we can on the input by using different filters andwe get distinct feature maps. Finally, we stake all the feature maps together and it isthefinaloutputoftheconvolution layer.

The size of the output neuron (the feature map) is controlled by three Hyperparameters: Depth, Stride, and padding. Theseparameters should be decided beforetheconvolutionoperation isperformed [29].

- **Depth**isthenumberoffiltersthatweuseontheconvolutionoperation. The larger the number of filters the stronger the model we produce, butthere is a risk of overfitting due to increased parameter count. During theconvolution operation, if we use three different filters, we will producethreedifferentfeaturemaps.Finally,thesefeaturemapsarestackedas2 Dmatrices,so,thedepthofthefeaturemapswouldbethree.
- **Stride**isthenumberofpixelsthatthefilterslidesontheinputvolumeat a time. When the stride is 1 the filter matrix slides 1 pixel on the inputvolume at a time. When the stride is 2 the filter jumps 2 pixels on the inputvolume at a time and so on. If the number of strides is higher the outputvolumewillbesmaller.
- **Padding**isaddingzerosintheinputvolumearoundborders.Itisconvenient to pad the input volume around borders with zeros. It helps tokeepmoreinformationaroundthebordersoftheinputandallowscontrolling the size of the feature map. Commonly filters with a size of 3,stride with 2, and padding with 1 are used Hyper parameters in CNN but,wecanchangetheseHyperparametersdependingontheinputvolumeweha ve[29].

To control the numberof free parameters in the convolution layer, there is asystematic method called parameter sharing. If one feature is useful to computesomespatialposition,itshouldalsobeusefulinanotherposition.Inotherwords,i fwe use the same filter (commonly called weights) in all parts of the input volume,the number of free parameters decreases. The neurons in the convolutional layersharetheirparametersandonlyconnecttosomepartsoftheinputvolume.Parameter sharing of resulting from convolution contributes to the translationinvariance of CNN, i.e. when the input volume has some specific centred structureand we want the CNN to learn different features in some other spatial location, inthiscase,wesimplysharetheparametersandcalllocallyconnectedlayer[29].

Finally, to make a single convolution layer we need to add the activation function(ReLU) and bias (b) to the output volume. The following figure (Figure 2. 6) [43]showsoneconvolutionlayerofCNNwiththeReLUactivationfunction.

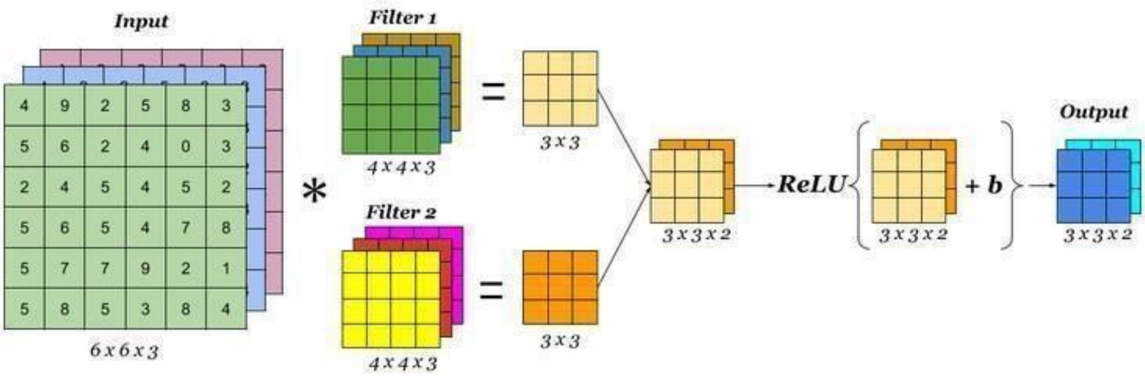


Figure2.5AnExampleofoneconvolutionlayerwithactivationfunction

B. PoolingLayer

To reduce the number of parameters, to extract dominant features in some spatiallocation, to progressively reduce the spatial size of the convolved feature, and tocontrol the problem of overfitting in the network we need to add pooling layer (alsocalled sub sampling or down sampling) in between some successive convolutionlayersinCNN[29].Thislayerhelpstoreducethecomputationpowerthatisrequi redto train the network. The pooling operation is performed by sliding the filter on theconvolvedfeature.

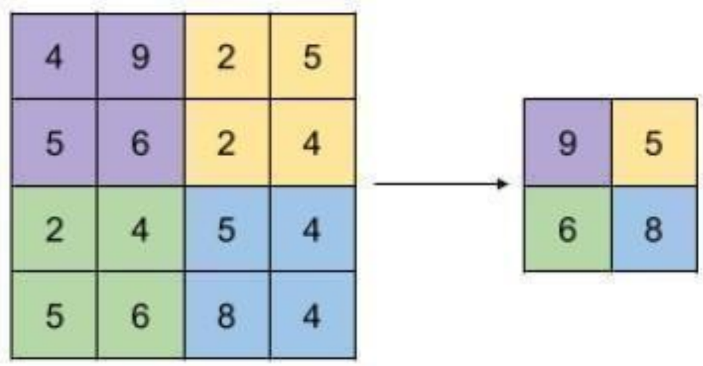


Figure2.7ExampleofMaxpooling

There are three types of polling: Max pooling, Averagepooling, and the lesscommonly used type, which is Sum, pooling. The max polling (Figure 2.6) [29] isthe most commonly used polling operation and its output is the maximum valuefrom the portion of the data covered by the filter. The average pooling returns theaverage of all the values from the data covered by the filter and finally, the sumpooling returns the sum of all the values from the portion of the data covered by thefilter. The max pooling performs de-noising along with dimensionality reductionbut average polling is only used for dimensionality reduction. Therefore, maxpollingisbetterthanaveragepooling.Thepoolingoperationisappliedinallofthedept h slices of the data after the convolution operation; the commonly used filter is2×2, and stride 2 but we can change it accordingly. For example, if we take thecommonly used 2×2 filter (as shown in Figure 2.7), for the max pooling, it returnsthemaximumvaluefromthefourvalues[27].

C. FullyConnectedLayer

The fully connected layer is the same as the traditional multilayer Perceptron that isdiscussed in Section 2.4.2.1 above. In a fully connected layer, every neuron in theprevious layer is connected to every neuron in the next layer. This layer accepts theoutput of the convolution or pooling layer that is high-level features of the inputvolume. These high level features are in the form of a 3D matrix but the fully connectedlayeracceptsa1Dvectorofnumbers.Therefore,weneedtoconvertthe3D volumeofdata into a 1D vector called flattening and that becomes the input to the fully connectedlayer. The flattened vector is given to the fullyconnected layer and it performsmathematical computationlikeanyANNandthecomputationisdiscussedinSection 2.4.2. Activation functions such as ReLU in the hidden layers are used to apply non-linearityintheselayers.Byusingthesigmoidactivationfunctionthelastlayers(output layer) of the fully connected layer perform classification (probabilities of inputs beingin a particular class) based on the training data. For example, in this study, the dataclassification will have two classes: Fraud and Nonfraud. In addition to classification,a fully connected layer is a better way of learning non-linear features of the outputreturnedfromconvolutionand poolinglayers.

2.6 ApplicationofDeepLearning

Financialfrauddetection

Deep learning is being successfully applied to financial fraud detection and anti-money laundering. "Deep anti-money laundering detection system can spot andrecognize relationships and similarities between data and, further down the road,learn to detect anomalies or classify and predict specific events". The solutionleveragesbothsupervisedlearningtechniques,suchastheclassificationofsuspicioustransactions,andunsupervisedlearning,e.g.anomalydetection[15].

Military

TheUnitedStatesDepartmentofDefenceapplieddeeplearningtotrainrobotsinnewtasksthrough observation[23],[26][31]

Customerrelationshipmanagement

Deep reinforcement learning has been used to approximate the value of possibledirect marketing actions, defined in terms of RFM variables. The estimated valuefunction was shown to have a natural interpretation as a customer lifetime value[23],[26][31].

Recommendationsystems

Recommendation systems have used deep learning to extract meaningful featuresfor a latent factor model for content-based music recommendations. Multi-viewdeep learning has been applied for learning user preferences from multiple domains.The model uses a hybrid collaborative and content-based approach and enhancesrecommendationsin multipletasks[23],[26][31]

Bioinformatics

An AutoEncoders ANNwas used in bioinformatics, to predict gene ontologyannotations and gene-function relationships. In medical informatics, deep learningwasusedtopredictsleepqualitybasedondatafromwearabledevicesand predictionsofhealthcomplicationsfromelectronichealthrecorddata.Deeplearninghas alsoshownefficacyinhealthcare [23],[26][31].

2.7 Relatedwork

Several authors have tried to study tax fraud detection, especially in developedcountries. There are many studies on tax fraud detection using data mining methods,machine learning, and deep learning technologies. Some of these are listed asfollows:Currently, in the area of the taxation system, the reduction of revenues andlossoftax(income)ismainlycausedbytaxfraud.Toreducetheselossesthereisaneed to develop a state of the art and automated method for tax fraud detection.Besides advancements in taxation, technologies are alreadydoing agreat jobincluding fraud detection using data processing techniques and in the last twodecades, the technology is getting faster and more accurate output. Basically, a lotof work has been done for tax fraud detection using data processing and machinelearning approaches.However, most of the studies conducted in the identificationof tax fraud are using the traditional data processing techniques and they follow acommonstep,whicharedataacquisition,datapre-processing,datafeatureextraction,andfinallyclassification[14,6,15,56].Differentclassificationtechniques are used in the literature such as Neural Network [57], support vectormachine (SVM), and some of the studies used both SVM and NN [58].There areno studies done in local flavour concerning business income tax Fraud detectionsince fraud natures are changing from time to time and the behaviour of frauds isdifferentfrom thedevelopedcountries.Mostresearchers usedclusteringandclassification techniques with k-Means and decision tree algorithms. In addition,most of the studies are implemented for specific domain areas. The main objectiveof this study is to apply Deep Learning to build a model that determines thefraudulent and non-fraudulent taxpayers to develop an effective tax collection bythe Ethiopian Ministry of Revenue. Therefore, to accomplish the tax audit operation,the authority needs to use the Deep Learning techniques to protect against fraud andimprove loyalty. In the following, we discuss literature in the area of tax frauddetection and classification, which are directly related to our study. The idea of thispaper is to increase the performance of tax data using the Bayesian network andParallelismtechniques.A parallelprocedureusedtheBayesiantechnique[21].

In the study “Fraud Detection on Bulk Tax Data Using Business Intelligence DataMining Tool”, the author of this paper, used a Mixed Methods Research (MMR),involving both Quantitative and Qualitative methodology and he used the outlieralgorithmmechanism [8].Anoutliercalls Datathatappear tohave differentcharacteristics than the rest of the population. The problem of outlier or anomalydetectionisoneofthemostfundamental issuesindatamining.

The weakness of this paper was that the dataset was very small which is not goodfordevelopingagoodperformancemodel.TheAuthorusestraditionalalgorithms.

In this paper the Author tried to extract high-risk taxpayers using the variance andthemean,andstandarddeviationthesuspiciousfinancialbehaviourisdetected,thejo b coefficient field is used, and high-risk occupations are identified and classified[7]. This paper provides an overview of the concept of Data Mining techniques anddifferent frauds in taxation. These techniques are DSS, fuzzy inference, and neuralnetworks. DSS is a specific class of computerized information systems that supportsbusinessandorganizationaldecision-making.Fuzzyinferenceistheactualprocessofmappingfromagiveninput toanoutputusingfuzzylogic[32].

Thispaperprovidesan explanationofanartificialneuralnetworkwhichisasingleneural network and focuses on small personal income taxpayers. The paper haslimitationswhenweseeitrelatedtoneuralnetworkconceptsandprinciples.

This paper focused on the machine learning approach to analyse tax fraud andfocuses on classification techniques rather than regression techniques. The paper haslimitations when we see it related to neural network concepts and principles. Theabove-listedpaperandotherrelatedpapersexplainsinTable2.2

Table2.2SummaryRelatedWorksonTaxFraud

No	Paper	Techniques	Noofdatasetused	References	Limitation
1	High-PerformanceImplementation of Tax FraudDetectionAlgorithm	Bayesiannetworkand Parallelismtechniques.	10028	[21]	It uses asingle datasetto train andtest thealgorithm.

2	FraudDetectionon Bulk TaxData UsingBusinessIntelligenceData MiningTool: A Case ofZambia RevenueAuthority	Outlierdetection	-	[8]	it does notclearly definethe data setthatisusedintheexperiment.
3	DetectingHigh-Risk TaxpayersUsing DataMiningTechniques	Linearregressionanalysisand SVM	33000	[7]	The paperdoes notconsider theimpact ofhumanfactors, suchas the qualityof the dataentry, on theaccuracy ofthealgorithm.
4	Application ofSoftComputingto Tax FraudDetection inSmall Businesses	Fuzzyinferenceand neuralnetwork.	-	[32]	The paperonly uses asingle datasetto train andtest thealgorithm.
5	On Big Data-based FraudDetectionMethod forFinancialStatementsof BusinessGroups	Clustering method(DecisionTrees,NeuralNetworks,Bayesian BeliefNetwork,K-Nearest Neighbour)	-	[33]	The paperdoes notconsider thecostofimplementingand using thealgorithm.
7	DetectingFinanci Using DataMiningTechniques:ADec Review from2004to2015	Surveypaper		[34]	The paperonly covers adecade ofresearch onfinancialfraud detection

					using dataminingtec hniques.
8	Characterizationa nd detection oftaxpayers withfalse invoicesusing dataminingtechn iques				The paperdoes notconsider theimpact ofchanges intax laws orregulation son theaccuracy ofthealgorit hm.
9	Financial FraudDetection withAnomalyFeatu re Detection	co- detectionfra mework	-	[35]	it does notclearly definethe data setthatisusedi ntheexperime nt.
10	Tax frauddetection through neuralnetwork s:Anapplicatio nusingasample of personalincom e taxpayers	NeuralNetwo rk	2,000,000	[36]	The paperdoes notconsider thecostofimpl ementingand using thealgorithm.
11	Machine Learning Approachfor Taxation Analysis using ClassificationTech niques	Bayes, Function, Meta	365	[37]	The paperdoes notevaluate theeffectiven essof theclassificati onalgorithm ona large- scaledataset.

Summary

Asmentionedintheprevioussections,thestudy showsthatdeeplearninghasbeenwidely used in the field of fraud detection, especially for tax systems, which isrelatedtobusinessinseveralways.Deeplearningtechniquesarealsoappliedfor

the detection and classification of different tax categories including business incometax but there is still a need to develop a more accurate and efficient model. As wesee in the related works (Section 2.6) all previously conducted, papers have someproblems, which we need to overcome in this study. For example, most of the papersusedtheirdatasetsfrominternetsearchesorpubliclyavailabledatabasessuchasin Kaggle that is recommended but the tax dataset in most of the previously conductedresearches are captured under controlled environments like in the laboratory setups.Therearemanylaboriouspre-processingstagessuchashandcraftedfeatureextraction, colour histogram, texture features, and shape features; most importantly,themethodsusedbypreviouslyconductedresearchworksarenotstateofthea rt,
i.e. most of the studies in the literature of tax fraud detection follow traditional taxdata processing techniques [13, 14, 15, 16]. In addition to this, the main point of thisstudy is that there is no tax dataprocessing using deep learning techniquesdesignedto detect or classify tax fraud detection so far. Hence, an accurate and efficient CNN-based model (avoids handcrafted feature extraction) for the detection of tax fraudbyusingtaxanalysis criteriaisdesignedanddeveloped.

CHAPTER THREEMETHODOLOGY(MATERIALS ANDMETHODS)

3.1 Introduction

This chapter focuses on the description of methodologies that are used in order toaccomplish the study including Flow of research, Methods of data preparation,softwareandhardwareconfigurationofthestudyused,andevaluationtechniques.

3.2 ResearchFlow

In this study, an experimental research method is followed in order to achieve theobjective of the study. As we can see in the process flow block diagram (Figure 3.1),thisstudyisconductedwiththreemainphases.Thefirstphaseincludesidentifyingthedomainoftheproblemwhichmeansunderstandingtheproblemandunderstanding the tax data. The second phase is about data preparation of the study.The third phase is the designed model is implemented with appropriate tools andmethods. The designed model is trained and tested with the appropriate data. Duringthe training of the model, the performance of the model is evaluated. After gettingtheoptimalmodelduringevaluation,themodelistested withtestdata.Finally,themodelispredictedby predictionmethods.

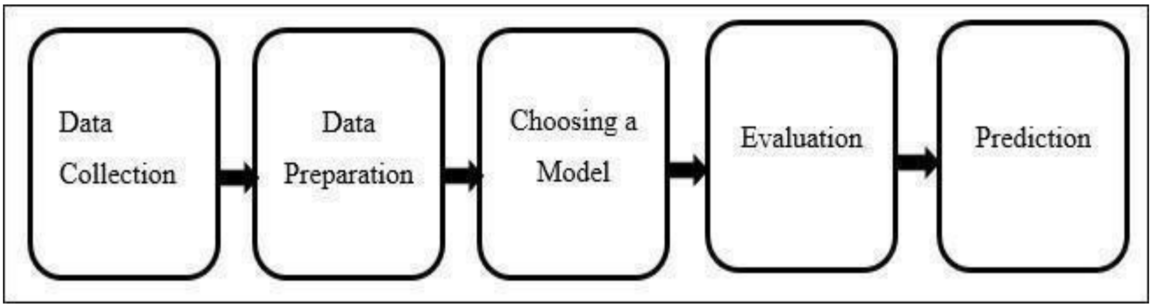


Figure3.1Researchflow

3.3 UnderstandingtheRevenuesDomain

In this study, the data collection methods will be employed to define the general workflow ofthe business income tax, and the domain experts and to understand the interaction of thedifferentDepartmentsintheMinistryof Revenues.

TheMinistryofRevenuesisthebodyresponsibleforcollectingrevenuefromcustomsdutiesanddomestictaxesinEthiopia.Inadditiontoraisingrevenue,MORisresponsibletoprotectsocietyfromtheadverseeffectsof trafficking[18].

MOR headquarters is in Addis Ababa which is led by a minister-level who reportsto the Prime Minister and is assisted by different offices or branches, namelyInternalAudit,TaxTransformationSecretaryOffice,CustomCommission,Institution Power and Support Branch, Minister Secretary Office, Tax OperationOffice, National Lottery and Tax Compliance & Risk Management Directoratebranch.

In MOR, thirty branch offices are available in Ethiopia, which comprise 22 CustomsControl stations, 50 Checkpoints, and 153 Tax Centres. Tax Centre means a taxcollection station administered under a branch office and located approximatelytaxpayers. This study understands the revenue collection task based on the SIGTASsystem.TheStandardIntegratedGovernmentTaxAdministrationSystem(SIGTAS) is the computer system that enables MORE taxes to be administered. Thesystem allows MOR to administer all aspects of most domestic taxes, includingRegistration, Assessment, Cashing, and Auditing in one easy-to-use integratedsystem.Thesystem wasintroducedinDecember1997.Currently,operatesinboththe head office and branch offices. One of the main activities of the authority isauditing(riskanalysisisauditandinvestigationaudit)thetaxpayers'financialstatements andbalancesheets.Theauditprocessandprogramdevelopmentdirectorate is working with the Information Technology Management Directorateclosely. The audit process departments to audit taxpayers' data firstly have the risk -analysed data. Based on this, the study focused on the risk analysis process to detectthe taxpayers during auditing time based on the previous year financial statementandbalancesheet.

TaxriskanalysingProcessandtaxfraudinvestigation

Under the tax Compliance & Risk Management Directorate sector, the Tax riskanalysing process and tax fraud investigation are organized. Thus, directorates havethefollowingactivities:

- Tochangethetaxandauditpolicyandstrategyof theauthorityintopractice.
- Tocreatefunctionalsystemsimprovedtaxandauditactivities.
- Toperformaspecialinvestigationauditandtransfertheresulttothecriminalinvestigationdirectorate.

MinistryofRevenuesRiskSelectionCriteria

The Ministry of Revenues needs to plan several screens. A screen needs to bedevised to create the different benchmarks used by the report. Another screen willbe devised to capture the details of the calculation of each benchmark while anotherwillbe needed to execute the calculation of the scores of the benchmarks tobe usedintheAuditRisk CriteriaReport.

The Report's objective is to assess the overall risk of a specific set of taxpayers byordering them according to risk factors. The report's intention does not necessarilytarget a specific tax type. The main purpose of the report is then to target who shouldbe audited. A proper audit case would follow. It can, also be used during an auditcasetoguide orsupporttheauditor.SinceschedulesCarethemaintaxesandtheirsum is used to define legally the annual turnover. The majority of the informationcaptured for audit risk criteria comes from them. In addition, information neededfor financial ratios comes from the financial statements and the tax declarations thataremandatory. Itisimportanttonotethatalthoughthereisataxselectioncriterionin the report, it is limited to schedules C. The MOR defines what mean risk andidentifiesthecriteriathatusedtocomputetheprofileoftaxpayerreport.

3.4 UnderstandingtheData

The first step in any Deep Learning problem is to collect more data to work with,analyse the data thoroughly, and understand the various parameters like Datasetcharacteristics,Attributecharacteristics,Numberof Instances and NumberofAttributes.

Afterunderstandingthedomainarea,we willhaveidentifiedthe data.Then,togetthe required data for this study an application letter addressed to the ITMD. TheData extracted from the Sources by an Authorised Database Administrator and thebe extracted Data scramble to achieve confidentiality. The data for this study willcompromise quantitative and qualitative data and levels of data measurement suchas nominal, ordinal, interval, and ratio toanalyse categorical dataand numeric data.At this stage, the data will be described briefly. The description includes a list ofattributes,theirrespectivevalues,and datatypes.Here,takingdataisnotenough fortrainingbecausedataintherealworldis composedofdifferentdataproblems [39]suchasinaccuratedata(missingdata),thepresentationofnoisydata(wrongdataandoutliers),and inconsistentdata.

Inthestudy,weanalysedthedataproblemstodevelopdeeplearningmodels,whichneedpuredatathatiseasy fortraining.

3.5 Datapreparation

Inthisstudy,taxpayers'recordonbusinessincometaxisusedasthemaininputtothe model. However, no publicly available database that contains taxpayers' datasetsworking on “Schedule C” tax schedule type that we can download and use for thetraining of the model. In this phase, firstly we will have focused on understandingthe revenue domain (business income tax type) and understanding the data asdescribed in Sections 3.3 and 3.4. Secondly, we will have to understand the pre-processingstepstopreparethetargetdataset.

3.5.1 DataPre-processing

In the real world, databases are highly susceptible to data problems. Due to this dataprocessing is the key issue. Data processing is the conversion of raw data into usefulinformation through a process [40], [41]. There are several methods and techniques,whichcanbeadoptedfortheprocessingofdata,dependingonthesoftware/hardwarecapability,timeconstraint,andavailabletechnology.Theseare:

- **Manual data processing** – In this type of data processing, the data areprocessed manually without the use of any electronic device or machine. Theprocess is slow and less reliable; it requires a large labour, and the chancesoferrorsbeing high[41].
- **Mechanicaldataprocessing**–Inthismethod,thedataareprocessedby using very simple devices like a typewriter or calculator. This method, whencompared to manual data processing, is more reliable and time-saving.However, theoutputcanstillbeverylimited[41].
- **Electronicdataprocessing**–Thismethodisfast,reliable,andaccurate. Computers are used to process data in electronic data processing. The labourrequired is minimal. Electronic data processing system, processing of a largeamountofdatawithhighaccuracyispossibletoimprovequalityand

maximize productivity.Therearethreestagesofprocesseddata.Inthefirststage, the collected data was inputted (domain expert and available data) intothe system (keyboarding or uploading). In the second stage, the data weremanipulatedandinthethirdstage, thedatawasprocessed[41].

In addition to the types of dataprocessing, the data pre-processing tools are appliedfor processing the data. There are different Data Pre-processing tools such as DataPre-processing in R, Data Pre-processing in Python, and Data Pre-processing inWeka[42].

For this study, we had selected the Electronic data processing type and data pre-processing tool in python [43], [44]. Data cleaning routines are applied to fill themissing values (with the mean value, and median value), smooth out noise (byremoving the record), and detect outliers (by removing or substituting with meanvalues, and median value) in the data. Feature selection consulting the domainexpertsandthedeeplearningusingpythonattributeselection-pre-processingtechniques (to reduce dimensionality) and by derivation of new attributes processedthe cleaned data. The result of these processes generates data sets for training andtesting.

StepsofDataPre-processing

In this study, the source data is organised by CSV file format. To convert the sourcedataintoacleardatasetweappliedthedeeplearningpre-processingsteps.Thefileloaded from the source by using numpy and pandas because we needed a data frameandNumpy forarray formatdatatopreparen_dimension matrixes.Thestepsare:

1. Getthedataset:

As described in section 3.2.1 we had to get the data from the Ministry of Revenue'soffice.

2. Datacleaning:

The data that we gathered was going to be messy, which may have inaccurateinformation or contain incomplete data like empty fields.In this phase, we had spentmore time to understand the data thoroughly, fill in the missing values, identifysmooth noisy data, identify or remove outliers, and resolve inconsistencies, andresolve redundancy caused by data integration. To solve the problems, we had usedmanualcheckuponexcelfilethenoisydata,medianstrategyforfillmissingvalue,and boxplotvisualisationtoidentifytheoutliervalue.

3. EncodingCategoricalData

Thisstudyuseddeeplearningneuralnetworksbutwhichrequirenumericinputandoutput variables.Therefore,wehadencodedthecategoricaldatatonumbersbeforedeveloping a model. There are many ways to encode categorical variables, althoughthethreemostcommonareasfollows:

- ❖ **IntegerEncoding**:anintegermappedtoeachuniquelabel.
- ❖ **OneHotEncoding**:abinaryvectormappedtoeachlabel.
- ❖ **Learned Embedding**: Where a distributed representation of the categoriesislearned.

Inthisstudy,wewillhaveusedtheintegerEncodingmethodaccordingtoourdata.

4. SplitthedatasetintoInputdataandlabeldata

Afterunderstandingthedatasetandencodingallcategorical features,wesplitthedatasetintoinputdata(featuredata)whichareindependentvariable (X),andlabeldata, whichisadependent variable(Y)usingtheSklearnlibrary.

5. Featurescaling

Featurescaling is amethod in datascience used tostandardize the rangeofindependent features in a dataset. Feature scaling would prevent the mentionedproblem and improve the overall performance of the model. Sample of featurescalingasdescribedin figure3.2[45].

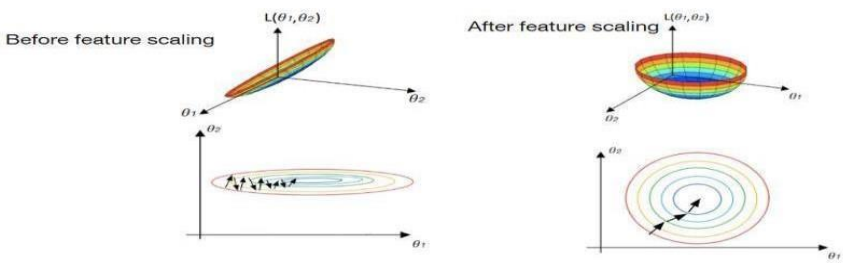


Figure3.2Featurescalingsample

In this case, the gradient descent can go straight towards the minimum of the lossfunction without any oscillation. In addition, it allows using a much higher learningrate, which reduces the overall training time of the model. Now that we have seenthebenefitsoffeaturescalinglet.

Recycling data is the process of making non-uniform attributes of a dataset uniform.Now, the question is when we would know whether a dataset is uniform or not.When thescale of attributesvaries widely thatcanbe ratherharmfulto ourpredictive model;wecallita nouniformdataset.Therescalingmethodisusefulinoptimizationalgorithmssuchas ingradientdescentwhichisdoneusingthe*MinMaxScaler*class,undersklearnlibraries.

Numericdatarepresentsdataintheformofscalarvalues.Thesescalarvalueshavea continuous range. This means there is an infinite amount of possible values.Integers and floating-point numbers are the most commonly used numeric datatypes. Numerical values are going to be the most frequent data types. Even thoughthey are already in a suitable format for calculations, the data may require some pre-processingsteps.Themainproblemwithnumericaldataisthedifferentscaleseachfeatur e holds [42]. In this study, we will use Normalisation, Standardization, andbinarizationtosolvethisproblem.

Normalisation

Normalisation simplyscales the values in the range [0 -1]. To apply iton adatasetwe have subtracted the minimum value from each feature and divide it with therange (max-min) as shown in the following equation as described in the *Equation*3.1[46]

Equation3-1NormalisationEquation $$xnew = \frac{x - xmin}{Xmax - xmin}$$

Standardisation

Standardisationontheotherhandtransformsdatatohaveazeromeanandone-unitstandard deviation. This can be achieved by the following equation as described inEquation[46]3.2

Equation3-2Standardisationequation $$xnew = \frac{x - \mu}{ð}$$

6. SplittheDatasetintoTrainandTestDatasets

The dataset is split into training data and testing the dataset after cleaning. We usethe training dataset to train our model and the test dataset to evaluate the trainedmodel,which is unseen during the training of the model. To evaluate better,we keptit completely separate and unique from the training data and test data. The validationsplit is used to assess the performance of the model, which is built during thetraining,andusedtofine-tunemodelparametersinordertoselectthebestperforming model. The literature recommends using the ratio of the training splitfrom 60% up to 90% of the total dataset and the rest for testing [10, 40]. In thisstudy, we have conducted the ratio **8:2**, which means **80%** of the dataset is fortraining and 20% of the dataset is for testing. From the training split, **20%** of thedataset is taken for the Testing Data set. Therefore, the training dataset contains18828 datasets, and the Testing dataset contains 4708 datasets. Since the two classes(fraudandNonfraud)haveanequalnumberofdatasetsineachcategory,thedatasetiss plitrandomlyintotrain,andtestedaccordingtotheratiostatedabove.Usinganequal number of datasets in each class for training and testing helps to avoid theproblem of overfitting because during the training updating of weights would not bebiased in one of the categories.Figure 3.3 Diagrammatical over all of the data pre-processingsteps:

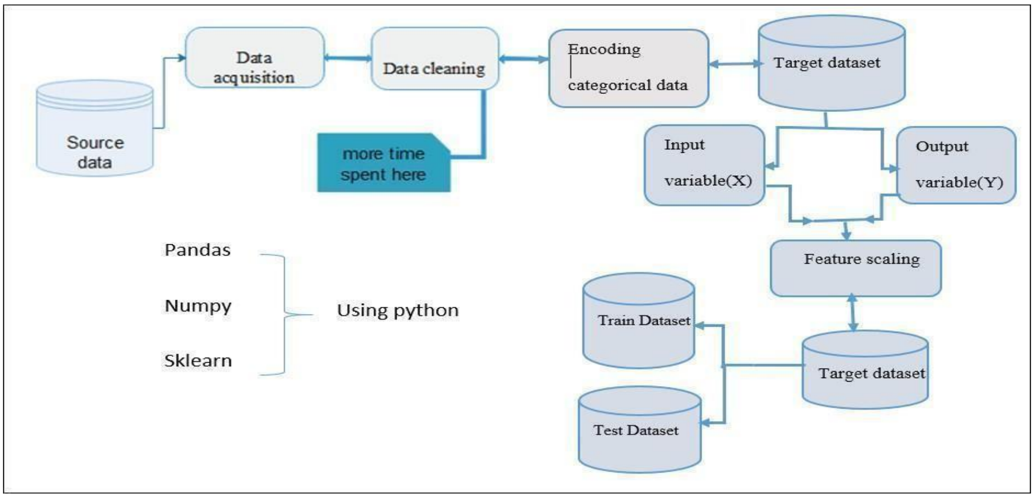


Figure3.3Diagrammaticaloverallofthedatapre-processingsteps

As shown in the above Figure 3. 3 To prepare the data, we will be using differentmachinelearninglibrarieslikepandas,NumPy,SklearnwithpythonProgramm inglanguage. As seen in Figure 3.3 The target dataset splits into a train and test dataset,whichisreadyforconstructing models.

3.6 SoftwareTools

Before selecting the tools, we have considered some criteria, which are helpful toselecttheappropriatesoftwaretoolswiththeircorrespondinglibraries.

Themaincriteriaare thechoiceofprogramminglanguage thatwillbe used toimplement the algorithm. The other criteria are to select tools with enough learningmaterialssuchasfreevideotutorials,andexistingexperience,andtheotheroneisthe toolsmustbeusedinmachineswithlimitedresources(likeCPUonly).Softwaretoolsthat we have used to implement the CNN algorithm are Python as a programminglanguagewithTensorFlowandKeraslibrariesonananacondaenvironment.

Anaconda is used for the implementation of the model and a free and open-sourcedistribution of the Python and R programming languages for data science andmachine learning-related applications that aims to simplify package managementand deployment. Installing the Anaconda environment, we got the Conda library,Jupyternotebooklibrary,pythonlibrary,andmorethanahundredlibraries[47].

JupyterNotebook

Open-sourcewebapplicationforinteractiveandexploratorycomputingandallowscreatingandsharingofdocumentsthatcontainlivecode,equations,visualizations,andexplanatorytext.ItisaplatformforData Scienceatscale[48].WehaveusedaJupyterNotebooktoimplementthecodingpart.Itiseasyandrunsinawebbrowser.

Numpy

NumPyisthefundamentallibraryforscientificcomputingwithPython.Numpyiscentred on a powerful N-dimensional array object; it also contains useful linearalgebra, Fouriertransforms,andrandomnumberfunctions[49].

Scikit-learn

Scikit-learnis an open-source librarywhich consists of various classification,regression,andclusteringalgorithms tosimplifytasks.Itis mainlyusedfornumericalandpredictiveanalysiswiththehelpofthePythonlanguage[49].

Pandas

Pandas are used for data manipulation, analysis, and cleaning. Python pandas arewell suited for different kinds of data, such as tabular data with heterogeneouslytyped columns, Ordered and unordered time series data, arbitrary matrix data withrow and column labels, unlabelled data, and any other form of observation orstatisticaldatasets.

SeabornandMatplotlib

Seaborn and Matplotlib are two of Python's most powerful visualisation libraries.Seabornusesfewersyntaxand hasstunningdefaultthemesandMatplotlibismoreeasily customizable through accessing the classes. Seaborn is an amazing pythonvisualisationlibrarybuiltontopofMatplotlib.

TensorFlow

TensorFlow is a software library or framework, designed by the Google team toimplement machine learning and deep learning concepts in the easiest manner. Itcombines thecomputational algebraof optimization techniques for easy calculationofmany mathematicalexpressions[50].

ToinstallTensorFlow,itisimportanttohave“Python”installedinoursystem[50].

Keras

Keras is a deep-learning framework that provides a convenient way to define andtrainalmostanykindofdeep-learningmodel.ItiswritteninPythonandcanberunon top of TensorFlow, or Theano. Keras is an open-source neural network librarywritten in Python. It is very simple to develop a model, user-friendly, and easilyextensible with Python. Keras layers can be added sequentially or in many differentcombinationsinaveryeasyway.Regardinghardware,youcanrunKerasonCPU sandGPUsandswitchbetweentheminaveryeasyway[51],[52].

The core data structure of Keras is the Model class. There are two types of built-inmodels available in Keras: sequential modelswhich are composed of a set of linearlayers[12], [42],[43], and models created with the functional API which enables usto define a more complex model, such as multi-output models and directed acyclicgraphswithsharedlayers[42],[43].

In this study, we will follow the Keras model lifecycle (Model creation, Configurethe model, Training the model, and evaluation of testing data or prediction on newdata)[51].

AdditionalSoftwaretools

- EdrawMax:todesigndifferentDiagramsnecessaryforthestudy.
- MSWord2016**:**fordocumentationpreparationofthestudy.Thereasonwhy,isits compatibilitywithvariousplatformsanditiseasytousefeatures.
- Microsoftexcels2016**:**tohandlethedatasetandtocomputetechnicalissues
- MicrosoftPowerPoint2016**:**ForPresentation
- Webbrowser:torunthepythoncodeusingJupyternotebook
- Mandalaysoftware**:**Itisafree,opensource,whichisareferencemanagementtool.We haveselecteditforpreparingthereferencepartofthestudy.

3.7 Hardwaretools

To implement the CNN algorithm with the selected software tools a very slowmachine with CPU Intel(R) Core (TM) i5-4210u CPU @ 1.70GHz processor,memory 8 GB was used. No GPU, which is the most important hardware in deeplearning for computer, vision research and also we will have to use additionalhardwaretoolslikePrinter,andSecondarystoragedevice(externalharddisk, USBflashdisk).

3.8 EvaluationTechnique

After training our model, we need to know how the model generalises for never seenbeforedata.Thishelpsustosaythemodelisclassifyingwellwithnewdata,orthemodel is doing well only for trained data (memorising the data fed before) but notinnewdata(datathathasnotbeen seenbefore).Therefore,modelevaluationistheprocess of estimating the generalisation accuracy of the model with unseendata (inour case test data). It is not recommended to use training data for evaluating a modelbecause the model remembers all data samples, which are fed during training, whichpredicts correctly for all the data points in the training but not for data that has notbeen seen during the training.In this study, to check the performance of theproposed model we have used confusion matrix, Classification report and MetricsDerivedfromConfusionMatrix.

ConfusionMatrix

A confusion matrix summarises the number of instances predicted correctly orincorrectly by a classification model [53]. We used to evaluate the fraud detectionmodel; the standard metrics derived from the confusion matrix table are; Truepositive(TP),Truenegative(TN),Falsepositive(FP),andFalse-negative(FN).Inthis study, there are two classes (i.e. Fraud and Nonfraud) and therefore the matrixeshave a dimension of 2×2. For the Target dataset a confusion matrix is similarlydefinedin thatrowandcolumn 2×2matrix.

Table3.1TheconfusionMatrixforTaxFraud

ConfusionMatrix		**Predictedvalues**	
		Fraud	Nonfraud
Actual values	Fraud	TrueNegative(TN)	FalsePositive(FP)
	Nonfraud	False Negative (FN)	Truepositive(TP)

Basedonbelowprinciplesthenumbersoftruepositive(TP),falsenegative(FN),falsepositive(FP),andtruenegative(TN) calculatedforeachclass.

Terminologies[53]associatedwithConfusionmatrixis:

- **TruePositives(TP)-**Truepositivesarethecaseswhentheactualclassofthe data point was 1 (True) and the predicted is also 1 (True). From thecontext of this study, it defines the number of Non Fraud records that arecorrectlyidentified.
- **TrueNegatives(TN)-**Truenegativesarethecaseswhentheactualclass ofthedatapointwas0(False)andthepredictedisalso0(False).Fromthecontext of this study, it defines the number of Fraud records that arecorrectlyclassified.
- **FalsePositives(FP)-**Falsepositivesarethecases whentheactualclassof the data point was 0 (False) and the predicted is 1 (True). False is becausethemodelhaspredictedincorrectlyandpositivebecausetheclasspredictedwasapositiveone.
- **FalseNegatives(FN)-**Falsenegativesarethecases whentheactualclass oftheinstancewas1(True)andthepredictedis0(False).Falseisbecausethe model has predicted incorrectly and negative because the classpredictedwasanegativeone(0).Inthecontextofthisstudy,itdefinesthenumber of records that are incorrectly classified as legitimate activitieshoweverin facttheyareNonfraud.

MetricsDerivedfromConfusionMatrix

Belowarethecomputationmetricsoftheclassificationmodel,whicharederivedfromtheconfusionmatrix in Table3.1.

Accuracy:

To evaluate the performance of tax fraud detection in terms of correctness we willuse Accuracy. It measures the ability of a classifier in correctlyidentify all samples,no matter if it is positive or negative. It determines the proportion of correctlyclassified instances concerning the total number of instances of the test. We can saythat accuracy is the percentage of correctly classified instances over the total numberofinstancesinthetotaltestdataset[53].

Accuracy=TP+TN/(TP+TN+FP+FN)∗100 ....(1)

Recall:

The ratio of the total number of correctly classified positive examples divided to thetotal number of positive examples can be defined as Recall. High Recall indicatestheclassiscorrectlyrecognized(asmallnumberofFN).

Recall=TP/(TP+FN)*100...(2)

Precision:

To get the value of precision we divide the total number of correctly classifiedpositiveexamplesbythetotalnumberofpredictedpositiveexamples.HighPrecisionindicatesanexamplelabelledaspositiveisindeedpositive(asmallnumberofFP).

Precision=TP/(TP+FP)*100 ....(3)

F-measure: Since we have two measures (Precision and Recall), it helps to have ameasurement that represents both of them. We calculate an F-measure, which usesHarmonicMeaninplaceofArithmeticMeanasitpunishestheextremevaluesmore.The F-Measurewillalwaysbenearertothesmallervalueof PrecisionorRecall.

F1:2TP/(2TP+FP+FN) .... (4)

CHAPTERFOUR

DESIGNANDIMPLEMENTATIONS

4.1 Introduction

This chapter focuses on the design of the proposed model Architecture and Traincomponentsoftheproposedsystemweredescribedbriefly.

4.2 ProposedsystemArchitecture

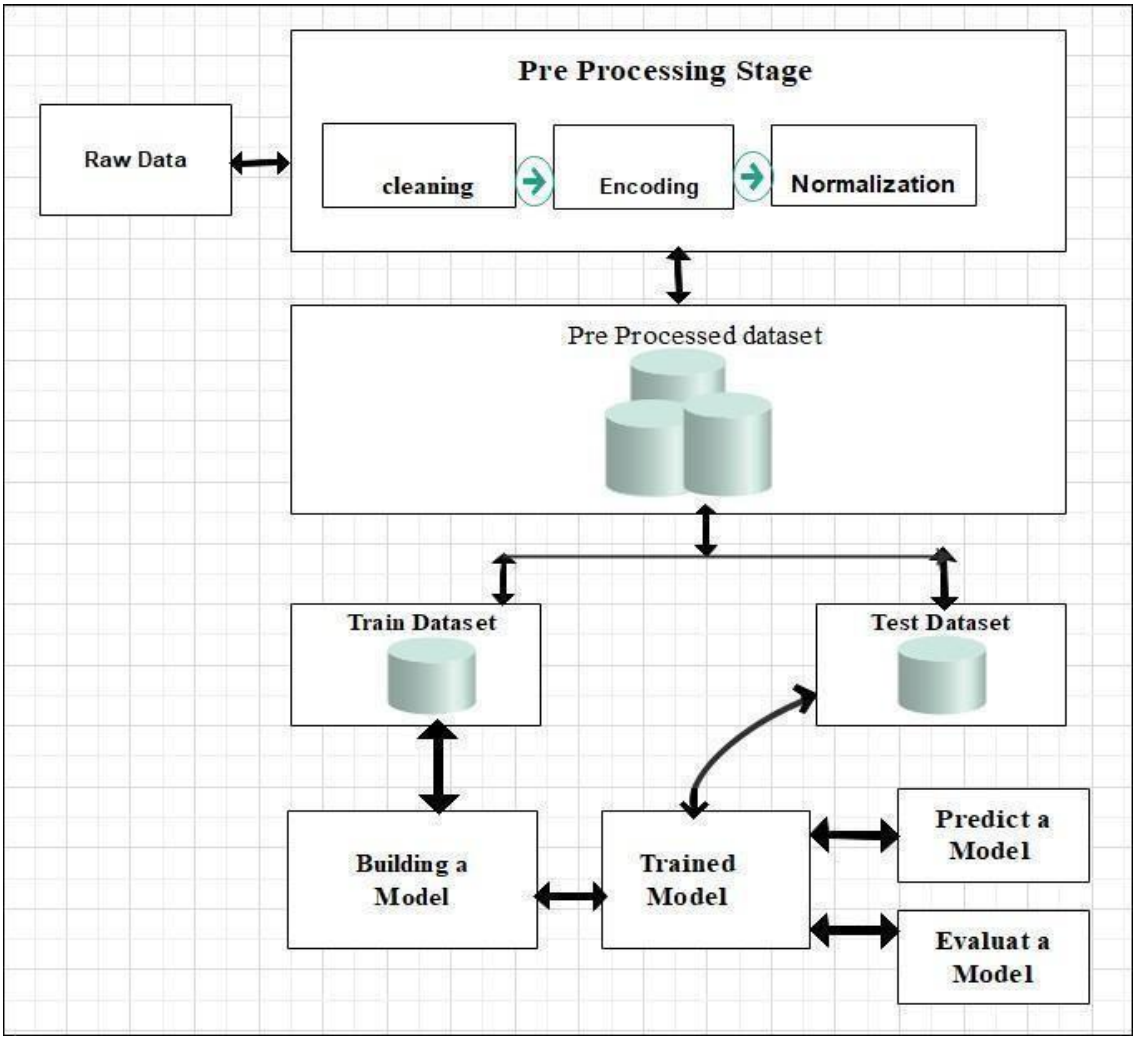


Figure4.1DiagrammaticOverallofResearchDesign

4.3 DescriptionofProposedSystem

4.3.1 InputData

Starting from the source database, several transformations are performed beforedesigning the target dataset as described before. When training a deep learningmodelthequalityofthetrainingdata determinesthequalityofthemodel[39].For this study,thedataisnotcleaninmostcasesasdescribedinChapter3Section3.5.In this stage, the raw data is organized based on the organization rules. First, the Taxrisk analyser developed the risk management criteria; design financial statementforms and balance sheet forms. Based on these forms the Taxpayers submitted thefinancial statement report to the auditor yearly or monthly. Some companies areusing Peachtree or Excel for day-to-day activities. Currently, the organisation doesnot use communication methods such as tax systems and web sites to facilitateinformation,reportexchange,andavoidphysicalinteraction.Theauthority'sa uditorsarecheckingthetaxpayers'incomeandexpendituresbasedonthetaxpayers' financial report. The authority rates informally the low/high annualincomesales,highgrossprofit/loss,highTotalExpenses,netincome/loss,refund able amount, and low total gross income as fraud suspicious claims. Theauditors rate lower gross profit/loss, net income/loss, tax due/refundable amount,non-operating income, low-profit income tax, low total expenses, and low totalgross income as Nonfraud suspicious claims. In addition to the above-mentionedcriteria, which were used by experts for judging whether a tax claim is fraudsuspicious or not, the type of tax/business and income class can also be employedfor the investigation of claims whether they are suspicious of fraud or not. Theprivate companies are also considered for showing fraud suspicious claims mostlybecause they may be having branches, sister companies and foreign companieswhile claims with government companies are mostly believed to be free of freak.All the informationwas gathered during claim processing and submitted to theheadof the auditor. The head assigns the auditor/s to investigate the case. After theinvestigation,theauditor/sreportstheresulttothehead.Basedontheinvestigationres ult the head makes a decision. The authority can take the case to the court ifnecessary. The central database of federal taxpayers is found in Addis Ababa aroundMexicoSquare.ThedatabaseismanagedbytheITMDdepartment.Afterstudying the database thoroughly,we havegone through thirty-five (35) important attributes.In Table 4.1, the total number of records is summarised based on the taxpayer'scategoriesorincomeclass.

Table4.1NumberofDataBasedonIncomeClass

Department	**Taxpayerscategory**	**Numberofrecords**
Information TechnologyManagementDirectorate(ITMD)	A	11300
Information TechnologyManagementDirectorate(ITMD)	B	5330
Information TechnologyManagementDirectorate(ITMD)	C	7300
Total		23930

Alreadytheoriginaldatawascollectedasdescribedinchapter3andChapter4butthe entire dataset was not taken directly to develop a model before eliminatingirrelevant and unnecessary data.Originally, in this study, there were 23,930 recordsandthirty-five(35)attributes asdescribed above.Herewehadanalysedtheinteraction of attributes to select the relevant data. We had used matrix correlationtechniques which are the basis for factor analysis, canonical correlation, and otherstatisticaltechniques thatreproducethestructureof therelationshipbetweenvariables or inputfeatures.A visualdisplayof the correlationmatrix of the selectedtaxdatasetisgivenin Fig.4.2.

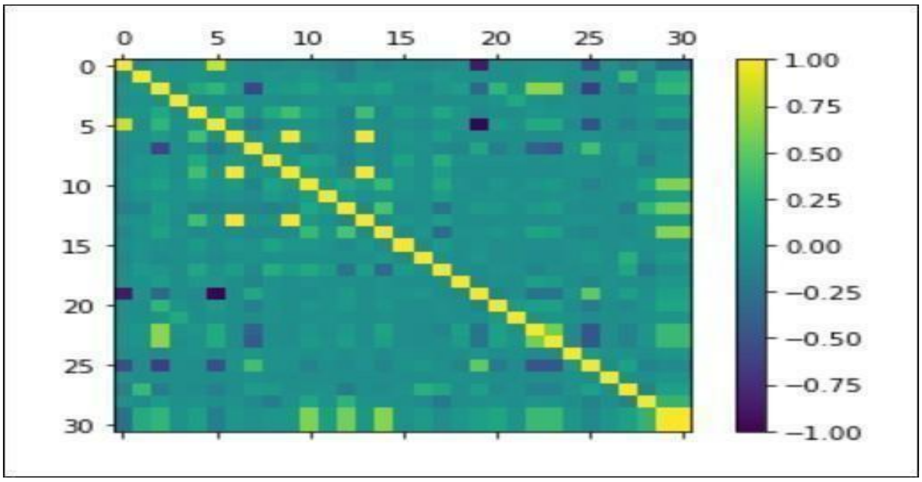


Figure4.2MatrixCorrelationbetweenFeaturesbeforePre-processingPhase

4.3.2 DataCleaning

Consequently, data cleaning has become necessary to improve the quality of dataand to improve the performance of accuracy.Removing the records that hadincomplete, noisy (invalid) data and filling in missing values under each column.As a result, the researcher used MS-Excel 2016 built-in functions like search andreplace, filtering, and auto-fill mechanisms, and Pandas library using Python toidentifyandfillmissingvaluesbasedonthehelpofexporters.

HandlingMissingValues

Missing values refer to the values for one or more attributesin datathat do not exist.In the real world, the Missing values in a dataset are common and it is a maliciousproblem.Ofcourse,this issuemustbe appropriatelyhandledbecause neuralnetwork models cannot work with this kind of data. From the total 23930 records,the maximum missing values are 4320 records,which contain 18% based onattributes percentage.Asshown below in Table 4.2outof the selected 35 attributesof they have registered with missing values. Accordingly, the researcher reacted totake appropriate action to clean the data, we had used different methods such asdroppingrowsandcolumnsmanuallywhenthereisnooptionandtheproblemwasknow n, we had used data imputation for numeric data by using the median strategy,andweusedStandardizationfeaturescalingtechniques.Themissingvaluesocc urredfortworeasons;thefirstoneduringdataentrytheclerkoftheITMDmadeamistakeandth eotherreasonwasthefinancialstatementsthatarenotfilledbythetaxpayers.

Table4.2HandlingMissingValues

Number	**Attributenameandtheirdatatype**	**Numberofmissingvalues**	**Datatypes**
1	**Annual_income_sales**	**2482**	**Numeric**
2	**Incomeclass**	**1047**	**String**
3	**Issueddate**	**4320**	**String**
4	**GrossProfit**	**284**	**Numeric**
5	**TotalExpense**	**176**	**Numeric**
6	**NetIncome**	**144**	**Numeric**

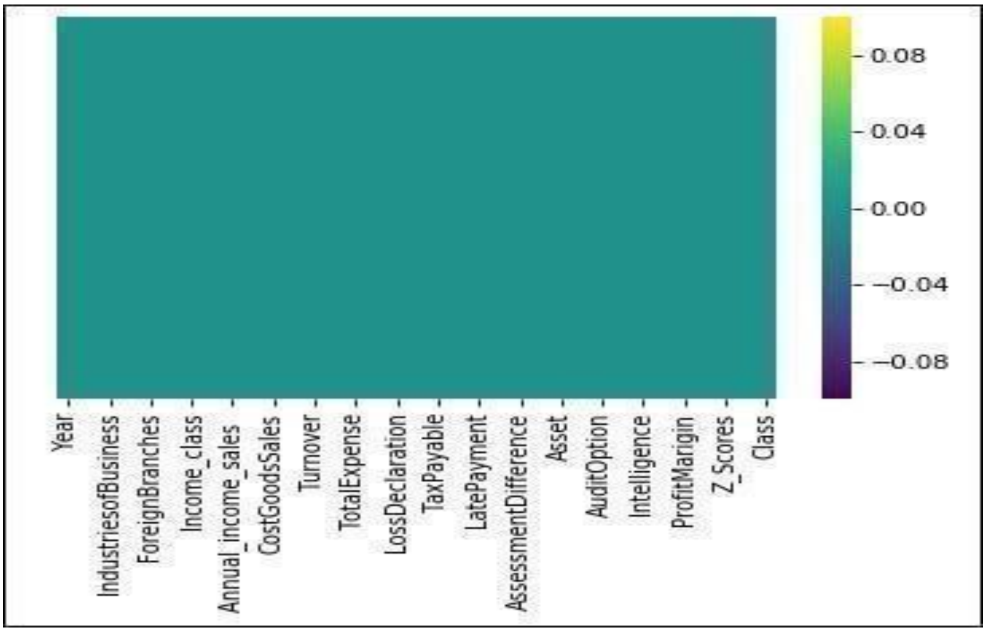
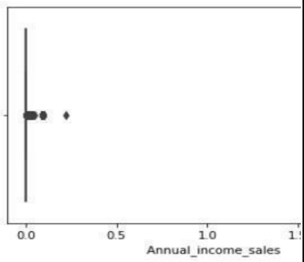
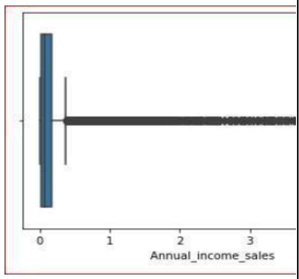
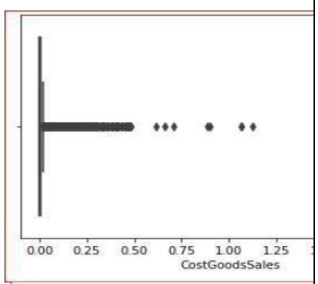
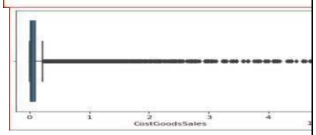


Figure4.3DiagrammaticviewofMissingValueHandling

HandlingOutliers

Inthisstudy,theresearcheridentifiedanddetectednoiseoroutliervaluefromthetax data by the help of domain experts, the identified outlier was corrected usingmanually,visualizationmethod,andstatisticaltechniqueslikeskews,median,etc. Accordingly,ourdata,weusedattributes**Annual_income_sales**and

Cost_of_Goods_Salesas asample todetect the outliervalue usingstatically andvisualisation techniques in addition to manually analysing as shown Table 4.5*Table4.3OutlierValueHandling*

Attribute	Methodtohandleoutliervalue		
	Statically(median)		Visualisation(boxplo t)
	Beforemedian	Aftermedian	
Annual Income Sales	**count 2.391900e+04 mean 1.198176e+07 std 1.629844e+08min 0.000000e+00 25%1.369998e+05 50% 5.656935e+0575% 1.701872e+06max 2.091295e+10**	**count 2.391900e+04mean 4.479205e+06 std 4.302383e+06 min 0.000000e+00 25%1.369998e+05 50% 5.656935e+0575% 8.780027e+06max 8.780027e+06**	 
Costof Goods Sales	**count 2.359200e+04 mean 1.306814e+06 std 3.781026e+06min 1.000000e+04 25% 1.643026e+05 50% 3.557456e+05 75% 1.004693e+06max 2.010132e+08**	**count 2.358100e+04mean 1.265035e+06 std 3.125836e+06 min 1.000000e+04 25% 1.642366e+05 50% 3.554619e+05 75% 1.002466e+06 max 4.700694e+07**	 

The below table 4.5 shown that the field **Annual_income_sales** contains outliervaluewhichmeansthemaxvaluebeforethemediantechniquewas2.091295e+10but after using a median technique the max value becomes 8.780027e+06 andusing Visualization technique the outlier value in the right side at the middle theboxindicatesthatvalueistheoutlier.Toremovetheoutliervalue,weusedtodropmani pulationaftersorting thedata.

4.3.3 EncodingCategoricalData

In our dataset, there are four categorical columns, which are Business Type,Business Group, Income Class, and Risk Status. For this study, we had convertedthesedataintoanumericdataformatusinglabelencodingasdescribedinChap ter 3.Theresultoflabel-encodeddataisdescribedintable4.4.

Table4.4Label-EncodingSample

Before Encoding		After Encoding
1	Stationary	Stationary --> 30
	Wholesale Trading	Wholesale Trading --> 32
2	Animal and Animal product trading	Animal and Animal product trading --> 4
3	Transportation and related service	Transportation and related service --> 31
4	Construction contractor	Construction contractor --> 13
5	Merchandise and Food grocery trading	Merchandise and Food grocery trading --> 22
6	Sewing	Sewing --> 29
7	Electronic and Electric ties	Electronic and Electric ties --> 14
8	Real estate	Real estate --> 26
9	Agriculture output, Hunting, Forestry and Fishing	Agriculture output, Hunting, Forestry and Fishing --> 3
10	Hotel and Restaurant	Hotel and Restaurant --> 18
11	Advertising	Advertising --> 2
12	supermarket trading	supermarket trading --> 39
13	printing service	printing service --> 37
14	Auctions	Auctions --> 1
15	wood and atena trading	wood and atena trading --> 41

SplitDatasetintotheInputFeaturesandtheLabel

Afterputtingtheexcelfileintoasingleexcelfileand handlingmissingandoutliervalueproblems,inthisstudywehadsplitourdatasetintoinputfe atureandlabel class.Splitthedatasetintotheinputfeaturesas(X)andthelabelas(Y)asStatedinTable4.5.

Table4.5FeatureSplitSample

Featureorvariable	Nameformachine	**Split**
Independentvariable	X	Allcolumnsexceptclasso ne
Dependentvariable	Y	Classlabelonly

4.3.4 FeatureScaling

As described in Chapter 3 in Section 3.5.1 the dataset is scaled up by many methods.For this study, we had used the **Normalisation** rescaling method. Based on thenormalisation formula, we used the median strategy to transform the data into therangeof[0,1]becausetherangeisfixedwhichisbettertoremovenegativeranges.Now we have seen our data in array format, which is easy to process for the machineasenlighten inAppendix E.

4.3.5 TargetDatasetDescription

In this study, we had split our dataset into a train dataset and test dataset. The trainingdatasetis80%oftheentiredatasetandtheremainingisthetest dataset.

There are different variables to split the data set into the train and test data set. Inthis study,we used these variables such as (X_train,X_test, Y_train, Y_test,train_test_split, X, Y, Test_size, and random state). From the total 23536-targetdataset,80% of the records,which are 18828, take as training datasetand 4708 takeastest dataset basedontheabovevariablesdescribedinTable4.6asbelow.

Table4.6TargetDatasetSplitting

InputFeatures(X)	TargetClass(Y)	%Split
X-training	Y-training	80%
X-test	Y-test	20%

4.3.6 BuildaProposedModel

Aftersplittingthetargetdataset,wecandevelopafrauddetectionmodelonbusinessincometaxusinga deeplearningalgorithm,whichisCNNthatisdescribed brieflyin Chapter2.As themain advantageof using the CNNalgorithm forthe detectionoftaxfraud,itismorerobustandautomatedthanclassicalmachinelearningalgorithms [21]. Inclassicalmachine learning algorithms, there is aneed todevelopdifferentalgorithmsfordifferentproblems. Therefore,MLusesmorehandcraftedalgorithms,butinCNNoncewedevelopedamodel forthedetectionofbusinessincometaxfraud.Thenwecanapplyforotherrelatedtaxscheduleslikebuildingrental income tax and employee income tax, so, which is easier to generalize [20].InthisstudytoimplementtheCNNalgorithm,eachrecordinthetaxdatasetusesa5x5 filterparameterandcontainsacentred,greyscalesdigit.Wehad installed thenecessarylibrariesasDescribedinChapter3.

As described in Appendix D the Target dataset contains 24 attributes includingtarget class and 23536 records, which is ready for building a model. Then we canloadthefileasamatrixofnumbersusingtheNumPyfunctionandasadata,frameusing the Pandas libraries. There are twenty -three input variables and one outputvariable (the last column).As described above in Section 4.3.2 most of the attributesare derived attributes. Therefore, we were using only twenty-three attributes that aremoresuspiciousforfraudbasedonexpertsandriskanalysiscriteria.OncetheCSVfile isloadedintomemory,wesplitthetargetdatasetintothetraindatasetandtest dataset as described in Section 4.3.5. The data was stored in a 2D array where thefirst dimension is rows and the second dimension is columns, e.g. [rows, columns]asReferredin AppendixD.

Before building the model, we had normalized the values of the dataset from [0,255] to [0.5, 0.5] to make the network easier to train (using smaller, centred valuesusuallyleadstobetter results).Wealsoreshapeor resizeeachrecordfrom(5,5)to(5, 5, 1) because Keras requires the third dimension.As we had described in Chapter3, Every Keras model is either built using the Sequential class, or the functionalModel class. In this study, we used the simpler sequential model since the CNN is alinear stack of layers and the sequential constructor takes an array of Keras layers.As discussed before, for conducting this study, the python version 3.7 software isused. Here we used all CNN layers and CNN parameters to improve the network.Wedidmoreexperimentsusingnetworkdepthbyaddingandremovingconvolutional layers, using dropout to prevent overfitting, and using fully connectedlayers for classification as we specified in Appendix D. We had used the lossfunction (binary cross-entropy) to evaluate a set of weights, Adam gradient basedoptimizertosearchthroughdifferentweightsforthenetworkbecauseitautomatically tunes itself and gives good results in a wide range of problems, andaccuracy metric to collect and report during training since a classification.Duringtrainingamodel,forthisstudywehadappliedthetraining data(recordsorX_trainandlabelsorY_train),numberofepochs(iterationsovertheentire dataset)totrainfor, and test data that is used during training to periodically measure the network'sperformance against data it has not seen before. We had achieved **84.64%** testaccuracy after 300 epochs.We have trained our neural network on the entire datasetand we can evaluate the performance of the network on the same dataset. In thisstudy, we had to return a list of two values. The first value is the loss of the modelonthedataset andthesecondistheaccuracyofthemodel onthedataset.

4.4 TrainingComponentsoftheProposedModel

In this study, the architecture is deployed with limited hardware resources anddesigned foronlytwoclasses.Inordertofindanappropriatemodel,aCNNmodelis designed which will work pretty well in a small number of datasets with very lowcomputational resources like CPU and GPU. The proposed model has 4 convolutionlayers,fullyconnectedlayers,ReLUinthehiddenlayersisincludedasanactivati on

function to add nonlinearity during the training of the network, and dropout isincluded afterthefirst twofullyconnected layers to prevent theproblem ofoverfitting as described on Figure 4.4. The proposed model descriptions (modelsummary)aredescribedinAppendixF.

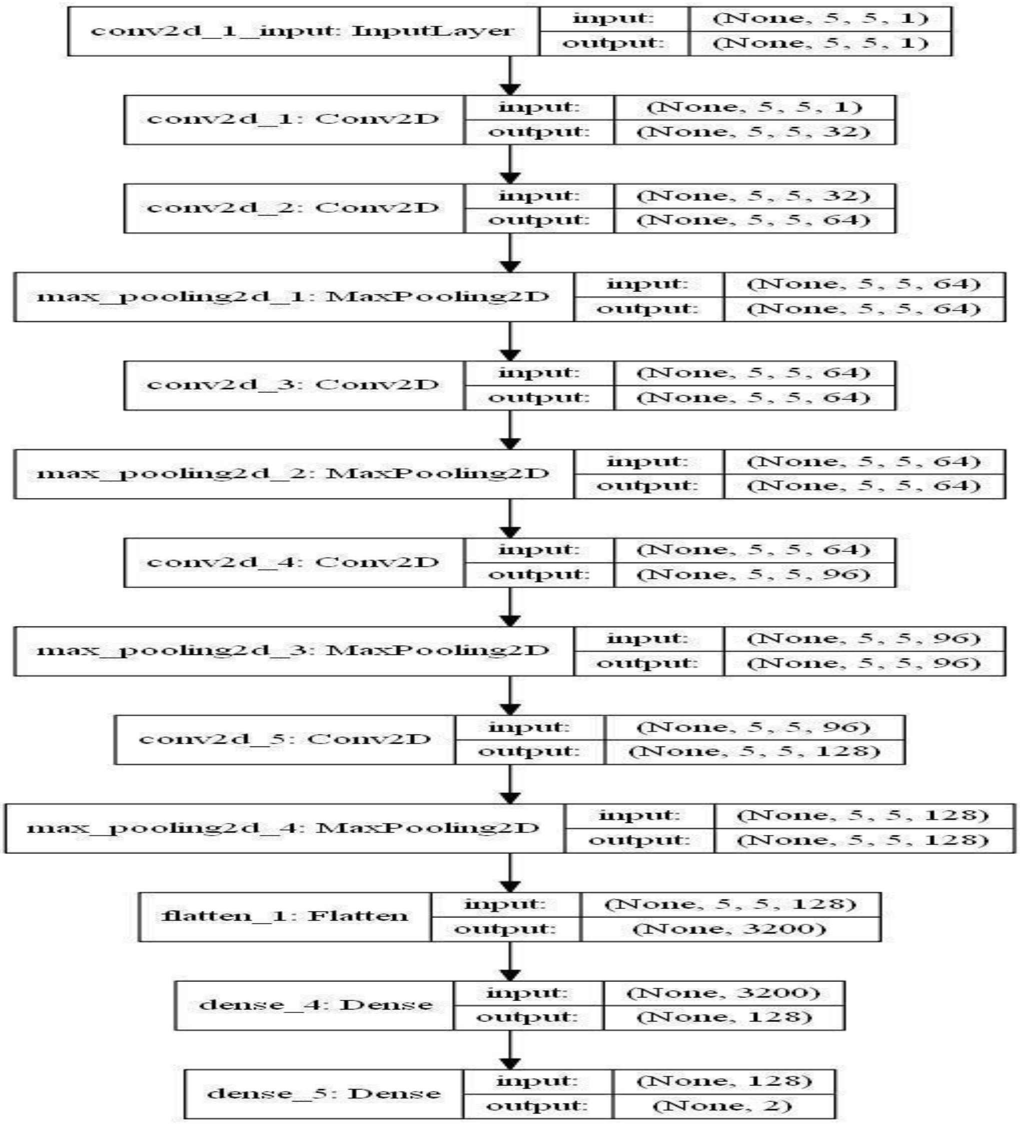


Figure4.4SVGArchitectureoftheProposedMode

4.5 EvaluationandPredictionStage

Afterthemodelisfit,predictionsaremade forallinthedataset,andtheinputrowsandpredictedclassvalueisprintedandcomparedto theexpectedclassvalue.

We have used a Softmax activation function on the output layer, so the probabilityof the prediction is in the range between zero and one. In this study, classificationaccuracy metrics are used which is a recommended technique for classificationproblems and when all the classes of the dataset have the same number of samples[21]. In this technique, the dataset is divided into training, validation, and testingdataset. During the training, we can feed the validation split to the model to getperformance metrics.Themodelreturnstheaccuracyandlossoftrainingdata,andthe accuracy and loss of validation data, which are training accuracy, validationaccuracy, training loss, and validation loss. Therefore, we can plot loss and accuracygraphs with respect to epochs by using these metrics. Finally, the testing data(dataset that has not been used in either the training or validation sets) is given tothe trained model to test the performance of the model, then the model returnsaccuracyandlossofthetestingdatawhichisneverseenduringthetraining.

4.6 DataAnalysis

InthisstudyfirstthedatasetfileformatwaspreparedinExcelfile format,whichisin CSV format. After that, this file is fed to the machine using pandas and Numpy.The machines train the data and then we could analyse the feed data in differentways using Seaborn and matplotlib. Finally, from the train data we had developedthemodel,themodel hadbeensavedbyJSONfileextensionorformat.

CHAPTERFIVE

MODEL EXPERIMENTATION AND DISCUSSION ONRESULTS

5.1 ModelExperimentation

This chapter describes the experiments made based on the described procedure in theprevious Chapters. Accordingly, python programming language got a 65% share of theavailable ML tools (Weka, orange, Java, and R). This shows Python is the most popularmachine-learning tool currently as described in Previous Chapters. We had used PythonProgramming Language because of the great number of packages with sufficientlibraries and documentation is easily available [31] for DL tasks.In this study, we usedan income tax dataset, which was prepared by ITMD and we adopted the supervisedclassification techniques.To develop a model in this study, first, we had split thedataset into the train dataset and test dataset. Then, we were reshaping the train and testdataset to rescale and normalize easily for the Keras model. After that, we fed the scaledand normalized data to the machine to develop a model.The model is implementedusing Keras sequential model technique within the CNN algorithm as described in thePreviousChapter.Thesetechniqueshadbeenimplementedusingananacondaenvironment on python programming language tools within different libraries such asPandas, NumPy,MatPlot,Seaborn,Keras, and TensorFlow. The description andevaluation presented the performances of the classification models. The methods,techniques, and algorithms of deep learning technology that were briefly explained inChapters 3 and 4 were applied to accomplish the objective of, the study. For featureextraction convolutional layers with activation functions (ReLU) and max poolingcomponentofCNNalgorithmwehadused.

5.2 ExperimentDesign

Before starting this experimentation part, the researcher discusses with the experts. Thisdiscussionfocusedonassessingtheinfluentialfactorsforbeingataxpayer.Generally,the experts were discussing some of the most important features and the researcherpointed outtheimportantpointsraisedbytheexperts.

The features are Industry_of_business, SaleTurnoverRatio, Loss Declaration, TotalExpenses, Intelligence, Custom, Risk Status, Audit Option, Commencement, Taxpayable,Branches,SisterCompany,ForeignBranches,Asset,ProfitMargin,Late payment, Refund, Turnover, Assessment Difference, Liability, Last audit, Tax Holidayand No_of_Emp had given a very high weight by the professionals. Consequently, inthe experimentation part, the analysis and interpretation of the model depends on theseattributes.However,thisdoesnotmeanthattherestoftheattributeshavenoimportance, rather it is to note the weight given to these variables in the real world bythe experts. As the experts explained, if a taxpayer's financial statement report has thefollowingcharacteristicsashavingahigherprobabilitytobefraudsuspicious.

1. Intelligencevalueoftaxpayers,high.
2. Commencement, Sales_turnover_ratio, Industry_of_business(constructionConstructor,ImportandExport),Branches,ForeignBranches, andSisterCompaniesvaluesoftaxpayers, high.
3. Turnover, Total Expense, Loss Declaration, Assessment (tax) Difference,andProfitMarginvaluesoftaxpayers,LatePaymentvaluesoftaxpayers,high
4. TaxPayable,TaxHoliday,andProfitMarginvaluesoftaxpayers,high
5. No_of_Emp,Refund,Asset,Liability,LatePaymentvaluesoftaxpayers,high
6. Last Audit, Audit Opinion, Profit Margin, Custom, Risk status values oftaxpayers,high
7. Sales_turnover_ratio,low,Industry_of_business,Branches,ForeignBranches, Sister Companies, Commencement, Turnover, Total Expense,Loss Declaration, Tax Payable, Assessment (tax) Difference, no_of_Emp,Refund, Asset, Liability, Late Payment, Last Audit, Audit Opinion, TaxHoliday, Intelligence, Profit Margin, Custom, Risk status, all have highvaluedepictedin AppendixA.

On the other hand, if a taxpayer financial statement report has the followingcharacteristics as having a higher probability to be Non- fraud suspicious.Sales_turnover_ratio,Industry_of_business,Branches,ForeignBranches,SisterCompanies,Commencement,Turnover,TotalExpense,LossDeclaration, Tax Payable, Assessment (tax) Difference, no_of_Emp, Refund,Asset,Liability,LatePayment,LastAudit,AuditOpinion,TaxHoliday, Intelligence, Profit Margin, Custom, Risk status, all have high, medium and lowvalue.Allfinalselectedattributesusedasaninputfor theexperiment.

All experiments were performed in a computer with the configurations Intel(R)Core (TM) 2 CPU 2.16GHz, 16 GB RAM, and the operating system platform.A procedure or mechanism of how to test the model's quality and validity hadneeded to be set before the model was built. To perform the model buildingprocessofthisstudy,an18828trainingdatasetwasusedtotraintheclassification models. Classification models had implemented using Pythonwith common deep learning libraries (i.e. NumPy, Scikit-Learn, Pandas, andMatplotlib) that contain libraries for data pre-processing, classification, andvisualisationandCNNalgorithmsusingDLKeraslibrary.Oncetheclassification model is developed, the performance of the model is checked outusing the test data set. Percentage split test options are used for training andtesting the classification model. This testing dataset was prepared by simplerandomsamplingtechniquesfromthetargetdataset.

In this study, three types of Experiments have been using to build the Deeplearningmodelsasshown inTable5.1

Table5.1ExperimentUsedtoBuildModel

Experiments	Model type(Keras)	DescriptionAboutExperiments
Experiment1	Sequential	CNNWithoutActivationFunction
Experiment2	Sequential	ImplementationofCNNwithActivationFunction
Experiment3	Sequential	Implementation of CNN with regularisationPerformanceImprovement and

5.2.1 ExperimentationI

Keras-CNN Model Experimentation without Activation function (ReLU) BasedontheDLmodelframeworkthedatasetselection,andpre-processingtechniquesareappliedthenDLmodeltraininghas beenmade.Inthisexperiment,theresearcherappliedthesequentialKerasmodeltypeandfollowedtheCNNalgorithm without Activation function (ReLU). The experimentation of this modelwasdonebyemployingthepercentagesplitclassificationmodels.Basedonthesetuptheclassificationmodelhadbuiltandthe resultfoundfrom thismodelsummarisedin Table5.2.

Table5.2ClassificationaccuracyusingKeras-CNNmodelwithoutActivationFunction.

Model	Numberoftestdatabaseinstances	Correctlytclassifiedinstance	Incorrectly Classifiedinstance	Correctly classified (%)	incorrectclassification(%)
Keras-CNN	4708	3173	1535	67.4	32.6

As shown in the confusion matrix, the Keras-CNN experimentation withoutactivation function has classified 2875 dataset records correctly while 1833dataset records incorrectly. Thus, the Keras-CNN experimentation withoutactivation function the model scored an accuracy of 49.56%. This indicated thatthe Activation functions affect the performance of the model, which minimisedtheaccuracy aspresentedin AppendixH.

5.2.2 ExperimentationII

Keras-CNNModelExperimentationwithActivationFunction

Inthisexperiment,weappliedallparametersexceptregularisation.Inthisexperiment, the researcher applied the sequential Keras model type and followedthe CNN algorithm with the Activation function (ReLU) without regularisation. Theexperimentationofthismodelwasdonebyemployingthepercentagesplitclassificationmodels.

Basedonthesetuptheclassificationmodelwasbuiltandtheresultfoundfromthis

Table5.3ClassificationaccuracyusingKeras-CNNmodelwithActivationFunction

Model	Number oftest datasetinst ances	Correctly classified instance	Incorrectly Classifiedi nstance	Correctly classified (%)	incorrectclassifi cation(%)
Keras-CNN	4708	3959	749	84.09	15.91

Asshownintheconfusionmatrix,theKeras-CNNexperimentationwithactivationfunction has classified **3959** dataset records correctly while **749** dataset recordsincorrectly. Thus, the Keras-CNN experimentation with the activation functionmodel scored an accuracy of 84.09%. This indicated that the Activation functionsbetter than the first experiment, which scored high accuracy but there was overfittingasdescribedin theNextSection5.3.

5.2.3 ExperimentationIII

Keras-CNNModelExperimentationwithRegularisation(Dropout)andPerformanceImprovement

In this experiment, we applied all parameters the same as the experiment I and IIexceptregularisationandincreasedtheepochvalue.Inthisexperiment,theresearcher applied thesequentialKerasmodeltype thesame as the previousExperiments and followed the CNN algorithm with regularisation and PerformanceImprovement to remove the overfitting. We had used dropout regularisation. TheExperimentationofthismodelwasdonebyemployingthepercentagesplitclassificationmodels.Basedonthesetup,theclassificationmodelwasbuiltandtheresultfoundfrom thismodelissummarisedinTable5.4.

Table 5. 4 Classification accuracy using Keras-CNN model with regularisation andPerformanceImprovement

Model	Number oftest datasetinst ances	t Correctly classified instance	Incorrectly Classifiedi nstance	Correctly classified (%)	incorrectcla ssification(%)
Keras-CNN	4708	3974	734	84.64	15.35

Asshownintheconfusionmatrix,theKeras-CNNexperimentationwithregularisation and performance improvement had classified 3974 dataset recordscorrectlywhile734datasetrecordsincorrectly.Thus,theKeras-CNNexperimentation with regularisation and performance improvement of the modelscoredanaccuracyof84.64%.Thisindicatedthattheregularisationandperformance improvement affects the loss value to evaluate the model,whichminimised the validation loss value as compared with Experiments I and II. In thisexperiment, the overfitting problem was solved by increasing the epoch's value,which is the iteration time greater than the previous experiments of the model,scoring an accuracy of **84.64%.** This indicated that the dropout and epochs minimisetheoverfittingofthemodelbecausetrainmorethantherestexperimentations.

5.3 AnalysisandDiscussionofResults

In this section, we show the results obtained from different models. Comparingdifferent techniques and selecting the bestmodelfordeveloping tax fraud detectionis one of the objectives of this study. Detailed analysis of each model is made in thebelowsections.

5.3.1 AnalysisofExperimentationI

In this Experiment, the models were used Keras-CNN without Activation Function.The model had been tested using the test validation technique. With this test, weevaluatedtheperformanceofthemodelagainstactualdataset entries. Specifically,in Figures 5.1 and 5.2, we had shown a comparative graph of the performance of themodelsfromlossrateandaccuracyusingtwotargetclassescorrespondingto0and

1. The result shows that the model has scored less performance than the rest basedonthemetrics. AsdescribedtheconfusionmatrixinAppendixH.

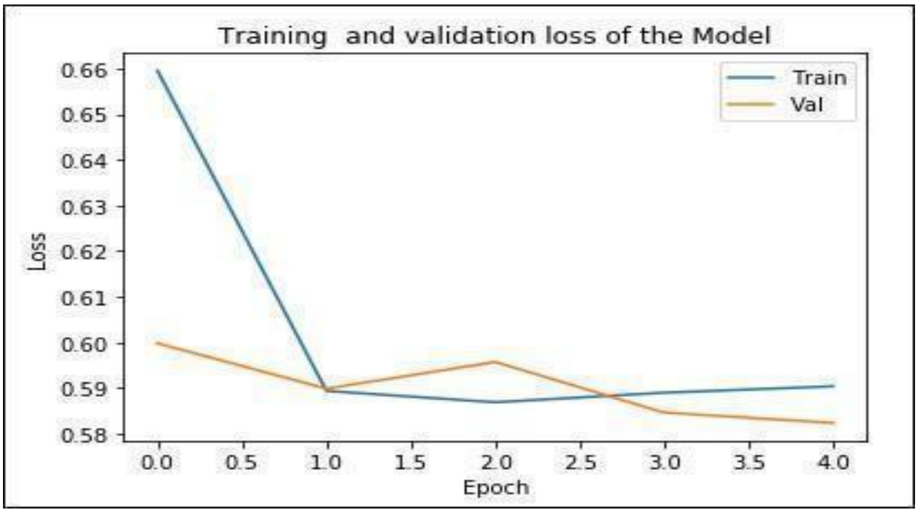


Figure5.1Loss-epochdiagramvisualisationonExperiment1

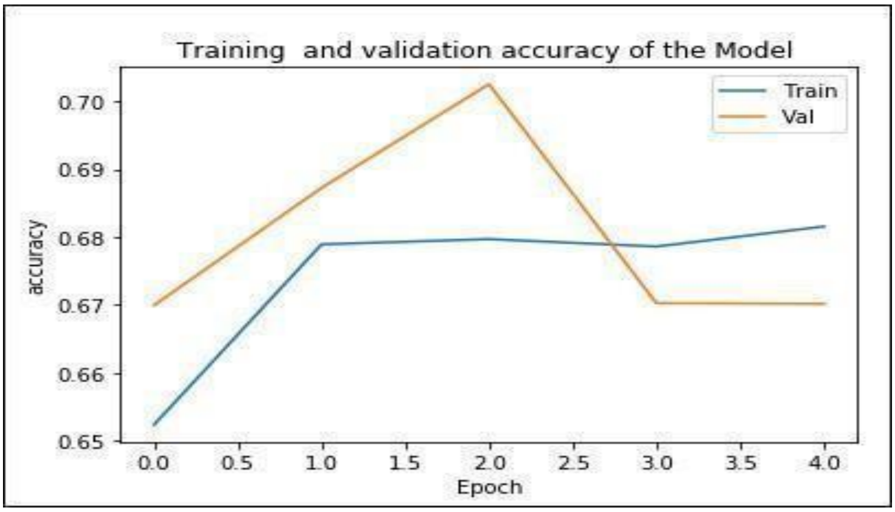


Figure5.2Accuracy-epochdiagramvisualizationonExperimentI

5.3.2 AnalysisofExperimentationII

In this experiment, the models used the Keras-CNN algorithm with ActivationFunction. The model had been tested using the test validation technique. With thistest,wehadevaluatedtheperformanceofthemodelsagainstactualdatasetentries.Specifically, in Figures 5.3 and 5.4, we had shown a comparative graph of theperformance of the models from loss rate and accuracy using two target classescorresponding to 0 and 1. As shown in Figure 5.3 and Figure 5.4 the model hadoverfitting on train and test data, which is the train, and test data needs overfittingremovalmethods.

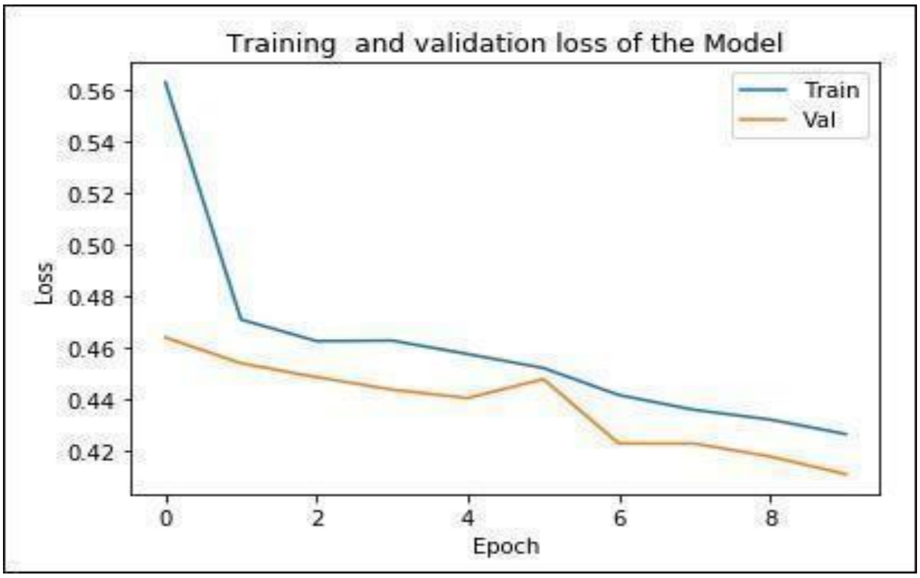


Figure5.3Loss-epochdiagramvisualisationonexperimentII

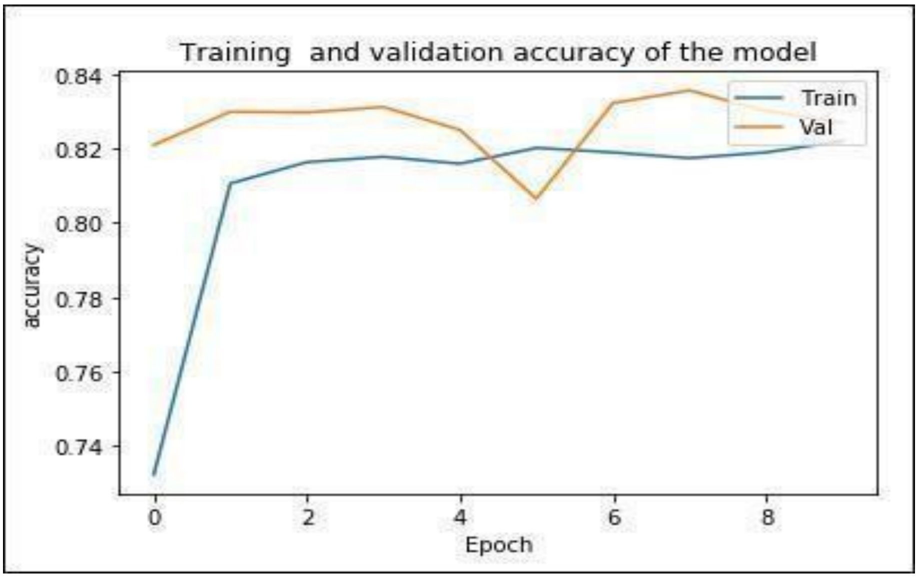


Figure5.4Accuracy-epochdiagramvisualizationonexperimentII

5.3.3 AnalysisofExperimentationIII

In this experiment, the models were used Keras-CNN within dropout regularisationand performance improvement value which was the epoch. The model had beentested using the test validation technique. With this test, we had evaluated theperformance of the models against actual dataset entries. Specifically, in Figures 5.5and 5.6, we had shown a comparative graph of the performance of the models fromlossrateandaccuracyusingtwotargetclasses correspondingto0and1.Theresultshowed that the model has scored higher performance than the rest based on themetrics and there is no overfitting on the training and test dataset because in thisexperiment,wehadusedtdropoutregularisationandwehadincreasedtheepoch value.This indicated that the train or iteration increased the performance of themodelandalsoincreasedindeeplearningneural networkconcept.

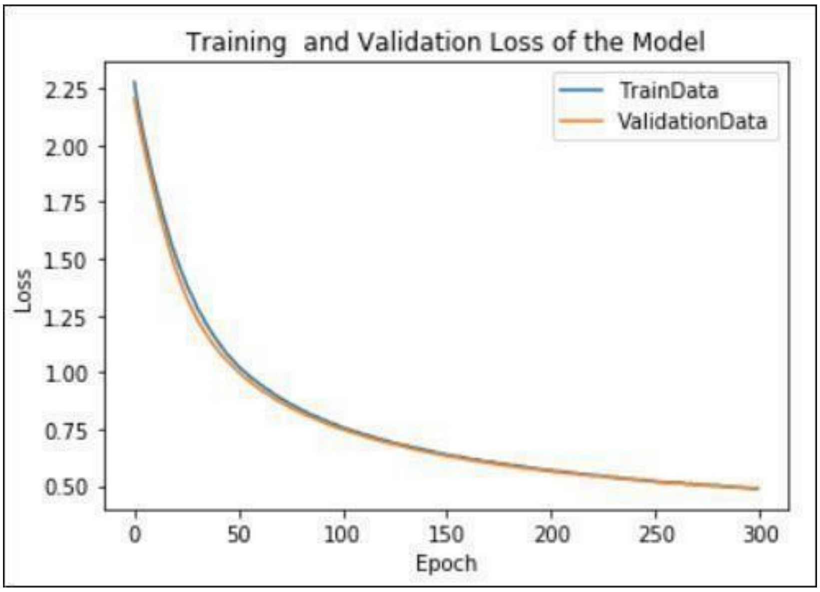


Figure5.5Loss-epochdiagramvisualizationonExperimentIII

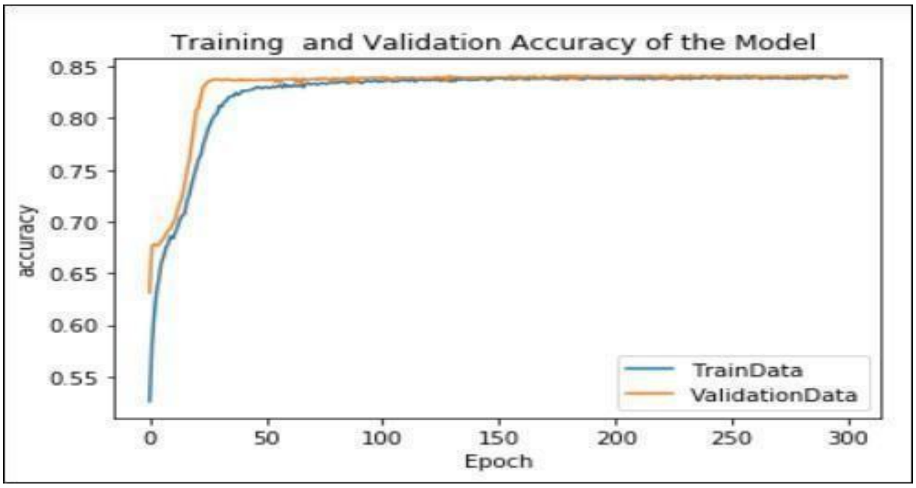


Figure5.6Accuracy-epochdiagramsvisualizationonExperimentIII

5.4 VisualisingtheProposedModel

In this study, we used Matplotlib and Seaborn libraries to create graphs such as LinePlots, Histograms, Three-dimensional plots, Steam plots, Bar charts, Pie charts,Tables, Scatter plots, based on the demand of the problem at hand. By using theselibraries, we had evaluated the Under fitting or Overfitting by Visualising thetraining loss vs. validation loss or training accuracy vs. validation accuracy over anumberofepochsisagoodwaytodetermineifthemodelhasbeensufficiently trained. We had adjusted the Hyper parameters: Hyper parameters such as thenumber of nodes per layer of the Neural Network and the number of layers in theNetworkcanmakeasignificantimpactontheperformanceoftheModel.Developing a model is not a success because the model should be checked out bytheperformanceevaluationmethod,which istheconfusionmatrix,andROCAUCcurve.

Confusionmatrix

A confusion matrix summarises the number of instances predicted correctly orincorrectly by a classification model as described in Appendix I.The developedmodel classified correctly **84.64%** of instances and classified incorrectly **15.35**% ofinstances. The different values of the Confusion matrix are explained in Figure 5.7basedonExperimentIII.

- TruePositive(TP)=2364;meaning2364positiveclassdatapointswerecorrectly classifiedbythemodel
- TrueNegative(TN)=1610;meaning1610negativeclassdatapointswerecorrectlyclassifiedbythemodel
- FalsePositive (FP)= 359;meaning 359 negativeclass datapointswereincorrectlyclassifiedasbelongingtothepositiveclassbythemodel
- FalseNegative(FN)=375;meaning375positiveclassdatapointswereincorrectlyclassifiedasbelongingtothenegativeclassbythemodel

Thisturnedouttobeadecentclassifierforourdatasetconsideringtherelativelylargernumberoftruepositiveandtruenegativevaluesas showninFigure5.7.

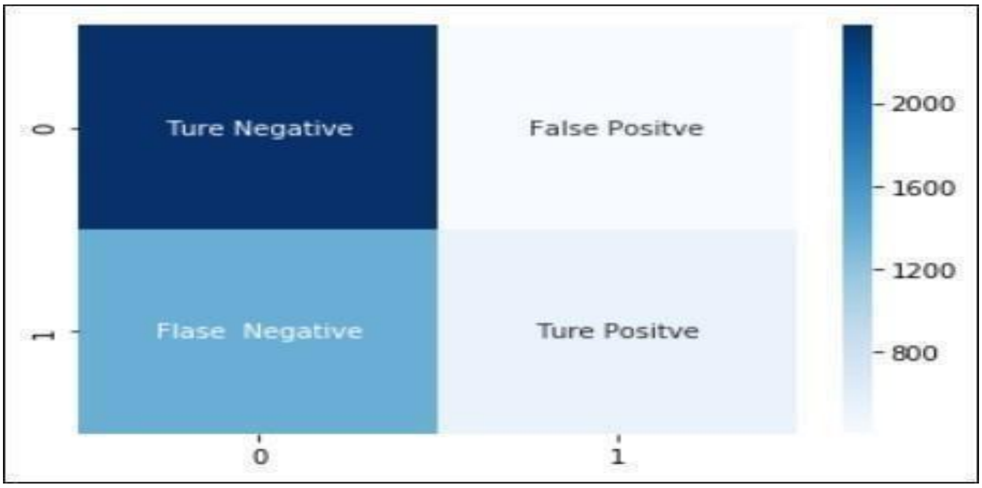


Figure5.7TN,TP,FN, FPConfusionMatrixDescription.

ROCCurveandAUCoftheModel

The ROC curve is good for viewing how the model behaves on different levels offalse positive rates [54] and the AUC (the Area under the Curve) are simple ways toview the results of a classification. For this study using the True Positive Rate (TPR)and False Positive Rate (FPR) formula, the result of ROC AUC in our experimentscored0.85or**85**%asshowninFigure5.8,whichisbasedontheAUCconcept.

Table5.5TheRequirementofROCCurve

ClassificationReport	**Precision**	**Recall**	**F1_score**	**Support**
Fraud	0.82	0.81	0.87	1923
NonFraud	0.86	0.87	0.87	2723
Sum	-	-	-	4708

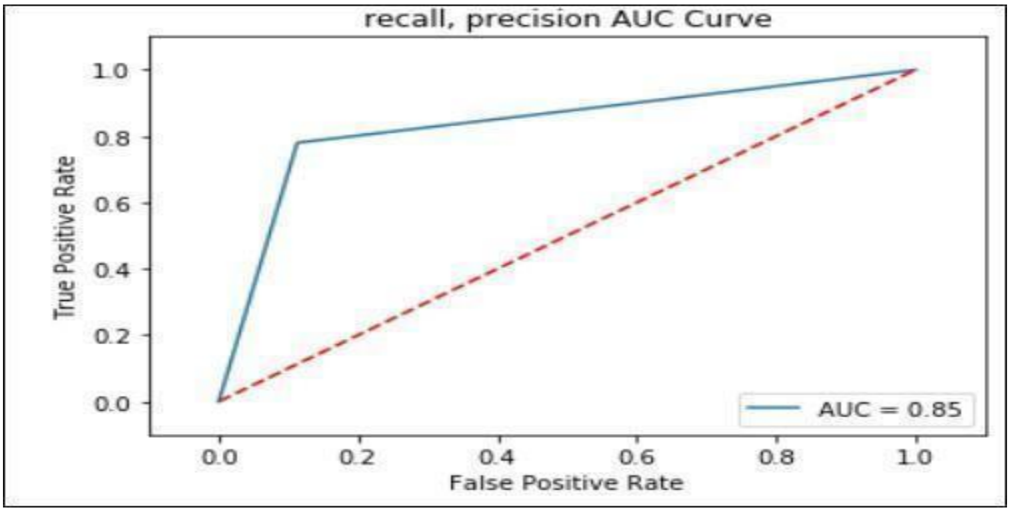


Figure5.8PrecisionRecallAUCcurveDescriptionTabl

e5.6SummaryoftheExperimentations

Experiment	#1	#2	#3
Accuracy(%)	67	84.09	**84.64**
Timetakentobuilda model(sec)	00:26.750706	3:03.141026	**0:03:58.697922**
Avg.Precision	0.68	0.82	**0.88**
Avg. Recall	0.62	0.80	**0.87**
Avg.ROC	62	84	**85**

CHAPTER SIXCONCLUSIONANDRECOMMENDATIONS

6.1 Conclusion

The technology of deep learning has increasingly become very popular and proved tobe relevantfor many sectors such as tax, insurance, airline, telecommunications,banking, andhealthcareindustries.Particularly in the taxsystem, Deep learningtechnology has been applied for fraud detection. Tax fraud is the most challengingproblem in the tax system. In this study, an attempt has been made to apply deeplearning technology to detect tax fraud. The machine learning process model hasfollowed while undertaking the experimentation. This process model embraces datacollection, preparation of the data, creating a Model, evaluation of the developed model,and checking the prediction value of the model. The data used in this study had gatheredfrom the Main database centres, which is ITMD. Once the data has been collected, thenthedatahasbeenpre-processedandpreparedina suitableformatforthedeeplearningtasks.Thisphasetookconsiderabletime.

The study was then conducted in two sub-phases, first, the phase of data pre-processingfollowed by the model-building phase. The initial data collected from MOR did notincorporate the target class for this study. The data pre-processing phase has beenconducted using pandas, NumPy, and Seaborn for segmenting the data into the targetClassesofFRAUDsuspiciousandNONFRAUDsuspicious.Bychangingtheparameters of the algorithm three different CNN Experiments have been conducted forgeneratingareasonablemodel.Themodelsfromthesethreeexperimentsareinterpreted and evaluated. Among the three Models, Experiment I has shown lessaccuracy value. The accuracy value of this experiment is 67.4% which is the Activationfunction value that affects the performance of the Model. The model developed withthe incremental epoch values and adding the regularisation (dropout) parameters haveshownabetterclassificationaccuracyof84.64%onthetrainingdataset.Thismodelisthen evaluated with a separate test dataset and scored an accuracy of 84.41% inclassifying new tax datasets as fraud and Nonfraud suspicious claims. This indicatedthat theiterationortraintimeincreasetheperformanceofthemodelalso

increasedandthedropoutregularisationavoidstheoverfittingvalueduringtrainingamodelA deeplearning-basedfrauddetectionmodelforthetaxsysteminEthiopia that used to achieve better performance. In general, the results from this study arevery promising. The study has shown that it is possible to identify those fraudsuspicious tax claims and suggest concrete solutions for detecting them, using deeplearningtechniques.Theproposedmodelisanalysedbasedonvariouskeyperformance indicators, which involves the statisticalparameters of precision,recall, overall accuracy, and F1-measure. The CNN observations of the performanceindicators are 84% (precision), 86% (recall), 84.64% (overall accuracy), and 87%(F1-measure). CNN is observed to be the best performer among all of the parametersexcept the precision, which is a least important factor among all four KPIs. Thisshows the efficiency of CNN classification with unnecessary feature correction. Inthe future, the fraud detection model on tax data can be improved further by usingdeeplearningwithconvolutional featuresuptomultiplelevels.

6.2 Recommendations

Thisstudyismainlyconductedforacademicpurposes.However,theresultsofthisstudy are found promising to address the practical problems of tax fraud. This studywork can contribute a lot towards a comprehensive study in this area in the future,in the context of Ethiopia. The results of this study have also shown that DeepLearning technologies particularly the CNN techniques in the Keras platform arewell applicable in the efforts of tax fraud detection. Hence, based on the findings ofthisstudy,thefollowingrecommendationsareforwarded.

Themodel-buildingprocessinthisinvestigationwascarriedoutintwosub-phases.

Fordatapre-processingtheresearcherusesthedataprocessingtoolsinPythonProgramming language with in whereas for classification CNN algorithm. However,the results were encouraging, but we were using only the CNN algorithm. Therefore,Further investigation needs to be done using other deep learning techniques such asRNN and AutoEncoders. In a work, only a limited number of all attributes are availablewith their values in the database of the authority. There are inconsistencies and missingvalues in the database. There is no data related to the number of withholding in thefirms, the total VAT, and TOT Since data is the most important component in Deeplearning study, the authority has to design a data warehouse where operational and non-operationaldatacan bekept.

- Inthis study,wedidnotconsiderindirecttaxtypes.Futureresearchcanbeconductedon thesetaxation systems.
- Frauddoesnotonlyoccurintaxcollection,butitcanalsooccurwithintheauthority of experts, auditors, and other staff. These can also be taken asanotherareaforfurtherresearch.
- WerecommendedthatforInformationexchange,andreportingpurposes different communication methods such as websites, applications, and othersystemswerebettertofacilitatetheactivitiesbetweenthetaxpayerandtheorganization.

REFERENCES

[1] M. Moges, "Perception on Factor Affecting the Efficiency of Income TaxAdministration in Large Taxpayer Branch Office of Ethiopia Revenueand Custom Authority," ADDIS ABABA UNIVERSITY COLLEGE,2017.

[2] M. Hanlon and S. Heitzman, "A Review of Tax Research," *Journal ofAccountingand Economics*.2010.

[3] D. Daba, "Tax Reforms and Tax Revenues Performance in Ethiopia,"2014.

[4] T. L. Gemechu, "The Ethiopian Income Tax System: Policy, Design AndPractice,"UniversityofAlabama,2014.

[5] FDRE,"FederalIncomeTaxProclamation,"2016.

[6] H. of peoples R. of the F. D. R. of Ethiopia, "Federal Tax AdministrationProclamation," *Federal Negarit Gazette of The Federal DemocraticRepublicofEthiopia*,ADDISABABA,2016.

[7] M. S. Rad and A. Shahbahrami, "Detecting High Risk Taxpayers usingData Mining Techniques," *2016 2nd Int. Conf. Signal Process. Intell.Syst.ICSPCS2016*,pp.14–15,2017.

[8] M. Mwanza and J. Phiri, "Fraud Detection on Bulk Tax Data UsingBusiness Intelligence Data Mining Tool: A Case of Zambia RevenueAuthority," *Ijarcce*,vol.5,no.3,pp.793–798,2016.

[9] S. Y. Huang, C. C. Lin, A. A. Chiu, and D. C. Yen, "Fraud DetectionusingFraudTriangleRiskFactors,"*Inf. Syst. Front.*,2017.

[10] O. for E. C.-O. and D. OECD, "Technology Tools to Tackle Tax EvasionandTax Fraud,"*Oecd*,2017.

[11] J. Schmidhuber, "Deep Learning in Neural Networks," *Neural Networks*.2015.

[12] Y.Lecun,Y.Bengio,andG.Hinton,"Deeplearning,"*Nature*.2015.

[13] R.E.Neapolitan, *NeuralNetworksandDeepLearning*.2018.

[14] J.Ahmad,H.Farman,andZ.Jan,“DeepLearningMethodsandApplications, ”2019.

[15] N.Ketkar,*DeepLearningwith Python*.2017.

[16] D. DeRoux, B. Pérez,A. Moreno, M. Del PilarVillamil, and C.Figueroa, “Tax Fraud Detection For Under-Reporting Declarations usingan Unsupervised Machine Learning Approach,” *Proc. ACM SIGKDDInt.Conf. Knowl.Discover.DataMin.*, pp. 215–222,2018.

[17] G.Lisi,“Taxmorale,taxcomplianceandtheoptimaltaxpolicy,”*Econ.Anal. Policy*,2015.

[18] B. Jean Bosco Harelimana *et al.*, “Effect of Tax Audit on RevenueCollectionin Rwanda,”2018.

[19] G. Q. Jira Jebessa, Fantahun Melles, Dieter Gagel, Ed., “Taxation inEthiopia,”in*DirectandIndirectTaxes-CategoriesofTaxpayersDeclarationofIncomeandAssessmentofTaxesTax*,2005.

[20] R.MENGISTU,“RevenueandTaxSysteminEthiopiaandtheEnforcement Problems in Addis Ababa City Administration,” AddisAbaba,2010.

[21] M. S. Rad and A. Shahbahrami, “High performance implementation oftax fraud detection algorithm,” *2015 Signal Process. Intell. Syst. Conf.Sp.2015*,pp.6–9,2016.

[22] S. ichi Amari, “Machine Learning,” in *Applied Mathematical Sciences(Switzerland)*,2016.

[23] M.Kubat,*AnIntroductiontoMachineLearning*.2017.

[24] R.D.Hof,“Deeplearning,”*Technol.Rev.*,2013.

[25] M.Kalash,M.Rochan,N.Mohammed,N.D.B.Bruce,Y.Wang,and F.Iqbal,“MalwareClassificationwithDeepConvolutionalNeuralNetworks,”in*Mobilityand Security*,2018.

[26] A.K.Tiwari,“Introductiontomachinelearning.”2017.

[27] A.Karpathy,“IntroductiontoConvolutionalNeuralNetworks,”2018.

[28] V.Mnih*etal.*,“Human-levelcontrolthroughdeepreinforcementlearning,”*Nature*,2015.

[29] D.A.Case*etal.*,“Amber 2017,”*Univ.California,SanFr.*,2017.

[30] F.Pedregosa*etal.*,“Scikit-learn:MachinelearninginPython,”*J.Mach.Learn.Res.*,2011.

[31] J.Ngiam,A.Khosla,M.Kim,J.Nam,H.Lee,andA.Y.Ng,“Multimodal Deep Learning,” in *Proceedings of the 28th InternationalConferenceonMachineLearning,ICML2011*,2011.

[32] C. Thang, P. Q. Toan, E. W. Cooper, and K. Kamei, “Application of SoftComputing to Tax Fraud Detection in Small Businesses,” *HUT-ICCE2006 First Int. Conf. Commun. Electron. Proc.*, vol. PART 1, pp. 402–407,2006.

[33] Y.J.ChenandC.H.Wu,“OnBigData-BasedFraudDetectionMethodforFinancial Statements of Business Groups,” *Proc. - 2017 6th IIAI Int. Congr. Adv.Appl.Informatics,IIAI-AAI2017*,pp.986–987,2017.

[34] M. AlBashrawiandM.Lowell,“DetectingFinancialFraudUsingDataMiningTechniques:a,”*J.DataSci.*,vol.14,no.3,pp.553–570,2016.

[35] X.C.DongxuHuang,DejunMu,LibinYang,“FinancialFraudDetection with Anomaly Feature Detection,” *IEEE*, vol. 3536, no. c,2018.

[36] C. P. López, M. J. D. Rodríguez, and S. de L. Santos, “Tax frauddetection through neural networks: An application using a sample ofpersonalincometaxpayers,”*Futur.Internet*,vol.11,no.4,2019.

[37] R. D. Lakshmi and N. Radha, “Machine Learning Approach for TaxationAnalysis using Classification Techniques,” *Int. J. Comput. Appl.*, vol. 12,no.10,pp.1–6,2011.

[38]ERCA, “የገቢ ግብር ነፃ መብቶች A ፈጻጸም መመሪያ,” *Federal Negarit Gazette*,2001.

[39] S.Zhang,C.Zhang,andQ.Yang,“DataPreparation forDataMining,” *Appl.Artif.Intell.*,2003.

[40] R. Sowmya and K. R. Suneetha, “Data Mining with Big Data,”in*Proceedingsof201711thInternationalConferenceonIntelligentSystemsand Control,ISCO2017*,2017.

[41] “Types of Data Processing,” *outsource2india*, 2016. [Online].Available:https://blog.outsource2india.com/types-of-data-processing.

[42] H.andKamber,“Data Preprocessing,”2006.

[43] R. Samarasinghe and M. Street, “Programming with Python,” *ESOFTComputerStudies*,no68.ESOFTComputerStudies,pallegama,2017.

[44] “HowToPrepareYourDatasetForMachineLearningInPython(1).”.

[45] M.Seeger,“Gaussianprocessesformachinelearning.,”*Internationaljournalofneuralsystems*.2004.

[46] D.M.vanRijmenam,“MachineLearning_HowtoPrepareforanAutomated Future,” *DataSeries*, 2019. [Online]. Available:https://medium.com/dataseries/7steps-to-machine-learning-how-to-prepare-for-an-automated-future78c7918cb35d.

[47] S.WestonandR.Bjornson,“IntroductiontoAnaconda,”2016.

[48] “JupyterNotebookDocumentation,”2020.

[49] “PandasDataFrametoNumPyArray.” .

[50] B.Pugh,“DeepLearningWithTensorFlow,”in*CMU*,2017.

[51] “KerasforBeginners_ImplementingaConvolutionalNeuralNetwork-victorzhou.”.

[52] A.BeamandF.Chollet,“TrainingDNNwithKeras,”*Arman.Vieira,BernardeteRibeiro*,2018.

[53] "ConfusionMatrixinMachineLearning,"*GeeksforGeeks*,2019.[Online]. Available:https://www.geeksforgeeks.org/confusion-matrix-machine-learning [Accessed:30-May-2020].

[54] S.Narkhede,"UnderstandingAUC-ROCCurve,"*TowardsDataScience*, 2018. [Online]. Available:https://towardsdatascience.com/understanding-auc-roccurve-68b2303cc9c5%22.[Accessed:26-Jun-2020].

APPENDICES

AppendixA:DescriptionofOriginalDatasetFeatures

No.	AttributeName	Data Type	Description
1	TIN	Number	Taxpayerstransactionidentificationnumberuniquely
2	Year	Number	Thetaxpayersregisteredyear
3	BusinessType	String	Thetaxpayersbusinessactivitiestheywork
4	Businessgroup	String	Taxpayerbusinesssector basedonasimilarcharacter
5	Annual_income_sales	Number	Taxpayersannualsaleincome
6	Income class (Category)	String	Categoryof thetaxpayerwhichisA,B,C
7	Sales_turnover_ratio	Number	Itdescribestheratiooftaxpayersthecurrentannualincomesaleandthepreviousannualincomesale
8	CGS (Cost of GoodSales)	Number	Whichindicatesthepurchasingofgoodsbeforesale
9	GP(GrossProfit)	Number	The different compute of CGS and annual income saleTurnover
10	Industry_of_business	Number	Describesbusinesstype
11	Branches	Number	NumberofBranchesexistence

12	ForeignBranches	Number	NumberofforeignBranchesexistence
13	Sister Companies	Number	Numberofsistercompaniesexistence
14	Commencement	Number	Taxpayersexistenceonbusiness
15	Turnover	Number	Sizeof Business/Turnover ofbusiness
16	TotalExpense	Number	These costs consist primarily of management fees andadditionalexpensessuchastradingfees,legalfees,auditorfeesandotheroperationalexpenses
17	ExpenseRatio	Number	Theratioofturnoverandaveragetotalexpense
18	Netincome	Number	NetincomethedifferencebetweenGpandExpenses.taxiscalculatedonit
19	LossDeclaration	Number	Numberof lossdeclarationfor serialyear
20	TaxToPay	Number	Amountoftaxtobepaid
21	TaxPayable	Number	Paidtaxdifferencebetweensucceedingyear
22	Assessment Difference	Number	Thedifference betweentheexistenceandthe currentone
23	No_of_Emp	Number	The numberofemployee workinginthecompany

24	Refund	Number	Differenceinrefundamountclaimed/Ataxcreditis notlimitedbytheamountofanindividual'staxliability.
25	Asset	Number	Average %ofthechangeinthetotalassetfrom thepreviousyear whenthedeviationisnegative
26	Liability	Number	Average%ofchangeintotalliabilityfromthepreviousyearwhenthedeviationispositive
27	Date	String	Thetaxpayerstaxauditeddate
28	LatePayment	Number	Taxpayer'sCompliance:-Number oflatepaymentsinthelasttwoyears
29	LastAudit	Number	ComparisontoDateofPreviousAudit
30	AuditOpinion	Number	Type ofauditoption
31	TaxHoliday	Number	Theavailabilityoftax
32	Intelligence	Number	3rdpartytaxinformationandintelligence
33	ProfitMargin	Number	Theratioofgrossprofitwithturnovertoidentifylowandhighprofit
34	Custom	Number	Customsprofile-basedoncomplianceleveloncustomsoperation(red,geen,yellow)
35	Riskstatus	String	Thestatusofriskwhichishigh,mediumandlevel

AppendixB:SampleSourceCode

```
from__future__importprint_function
importKeras
#fromkeras.datasetsimportmnistfrom
keras.utils import
to_categoricalfromkeras.modelsimpo
rtSequential
from keras.layers import Dense, Dropout,
Flattenfrom keras.layers import Conv2D,
MaxPooling2DfromkerasimportbackendasK
fromsklearn.model_selectionimport train_test_split
importnumpyasnpfrom
numpy import
array#readthefile
dataset=pd.read_csv("C://Users//hp//Desktop//weka//final//binarazationtrain.csv
",encoding='latin2')
#change into array
formatX=array(dataset.iloc[:,1:])X=np.resize(X,(X.
shape[0],img_rows,img_cols))Y=to_categorical(arr
ay(dataset.iloc[:,0]))
#splitthedatasetintotrainandtestdataset
(xtrain, Xtest, ytrain, Ytest) = train_test_split(X,Y,
test_size=0.2,
random_state=2)xtrain=xtrain.reshape((xtrain.shape[0],img_row
s,img_cols,1))
Xtest=Xtest.reshape((Xtest.shape[0],img_rows,img_cols,1))xtrai
n=xtrain.astype('float32')
Xtest =
Xtest.astype('float32')xtrain/=
255-0.5
Xtest/=255-
0.5print(xtrain.shape[0],'trainsample
s')print('x_train shape:',
xtrain.shape)print(Xtest.shape[0],
'test samples')print('x_train shape:',
Xtest.shape)model=Sequential()

model.add(Conv2D(64,
kernel_size=(1,1),activation='relu',kernel_regularize
r=regularizer s.l2(0.01),
kernel_initializer='he_normal',input_shap
e=input_shape))
model.add(Conv2D(64,(1,1),activation='relu',kernel_regularizer
=regularizers.l2(0.01)))model.add(MaxPooli
ng2D(pool_size=(1,
1)))model.add(Dropout(0.5))model.add(Flatt
en())
model.add(Dense(128,
activation='relu',kernel_regularizer=regularizers.l2(0.01)))
model.add(Dropout(0.5))
model.add (Dense (num_classes, activation='softmax',
kernel_regularizer=regularizers.l2(0.01)))
Model. Compile
(loss='binary_crossentropy',optimizer='Ada
```

m',

```
Metrics=
['accuracy'])exp4=model.fit(
xtrain,ytrain,batch_size=batc
h_size,epochs=epochs,
verbose=1,validation_data=(Xt
est,Ytest))
score=model.evaluate(Xtest,Ytest,verbose=0)print('Testlo
ss:',score[0])
print('Testaccuracy:',score[1])
# predict the first five test
datapredicts=np.round(model.predict(Xtes
t),0)#printthepredictionmodel
```

AppendixC:SampleCNNTrainedModel

```
Epoch 290/300
18828/18828 [==============================] - 1s 29us/step - loss: 0.5048 - accuracy: 0.8436 - val_loss: 0.5117 - val_accurac
y: 0.8405
Epoch 291/300
18828/18828 [==============================] - 1s 29us/step - loss: 0.5039 - accuracy: 0.8475 - val_loss: 0.5084 - val_accurac
y: 0.8388
Epoch 292/300
18828/18828 [==============================] - 1s 31us/step - loss: 0.5041 - accuracy: 0.8420 - val_loss: 0.5087 - val_accurac
y: 0.8422
Epoch 293/300
18828/18828 [==============================] - 1s 33us/step - loss: 0.5000 - accuracy: 0.8482 - val_loss: 0.5028 - val_accurac
y: 0.8396
Epoch 294/300
18828/18828 [==============================] - 1s 33us/step - loss: 0.5009 - accuracy: 0.8454 - val_loss: 0.5032 - val_accurac
y: 0.8437
Epoch 295/300
18828/18828 [==============================] - 1s 32us/step - loss: 0.4976 - accuracy: 0.8475 - val_loss: 0.4998 - val_accurac
y: 0.8424
Epoch 296/300
18828/18828 [==============================] - 1s 32us/step - loss: 0.5009 - accuracy: 0.8454 - val_loss: 0.5013 - val_accurac
y: 0.8437
Epoch 297/300
18828/18828 [==============================] - 1s 32us/step - loss: 0.4948 - accuracy: 0.8480 - val_loss: 0.4987 - val_accurac
y: 0.8424
Epoch 298/300
18828/18828 [==============================] - 1s 33us/step - loss: 0.4979 - accuracy: 0.8457 - val_loss: 0.4978 - val_accurac
y: 0.8437
Epoch 299/300
18828/18828 [==============================] - 1s 33us/step - loss: 0.4949 - accuracy: 0.8496 - val_loss: 0.4952 - val_accurac
y: 0.8407
Epoch 300/300
18828/18828 [==============================] - 1s 33us/step - loss: 0.4913 - accuracy: 0.8477 - val_loss: 0.4958 - val_accurac
y: 0.8441
```

AppendixD:TheResultofActualValuesandPredicted/ExpectedValues

[1.0,0.0, 1.0,1.0,0.0,0.0,0.0,1.0,1.0, 1.0,1.0, 0.0,1.0, 1.0,1.0, 0.0,1.0, 1.0,0.0, 1.0,0.0, 1.0, 0.0,0.0,0.0, 0.0,1.0,0.0, 0.0,1.0, 1.0,1.0,1.0, 0.0]=>1(expected1)

[1.0,0.0,1.0,1.0,0.0,0.0,0.0,1.0,1.0,1.0,1.0,0.0,1.0,1.0,1.0,0.0,1.0,1.0,0.0,1.0,0.0,1.0, 0.0,0.0, 0.0, 0.0, 1.0, 1.0, 0.0, 1.0, 1.0, 1.0, 1.0, 0.0]=>0(expected0)

[1.0,1.0,1.0,1.0,1.0,0.0,0.0,1.0,1.0,1.0,1.0,0.0,1.0,1.0,1.0,0.0,1.0,0.0,0.0,0.0,0.0,1.0, 0.0,0.0,0.0, 1.0,1.0,0.0, 0.0,1.0, 1.0,1.0,1.0, 0.0]=>0(expected0)

[1.0,0.0,1.0,1.0,0.0,0.0,0.0,1.0,1.0,1.0,1.0,0.0,1.0,1.0,1.0,0.0,1.0,1.0,0.0,1.0,0.0,1.0, 0.0,0.0,0.0, 0.0,1.0,1.0, 0.0,1.0, 1.0,1.0,1.0, 0.0]=>0(expected0)

[1.0,0.0, 1.0,1.0,0.0,0.0,0.0,1.0, 1.0,1.0, 1.0,1.0, 1.0,1.0,1.0, 0.0,1.0, 1.0,0.0, 1.0,0.0, 1.0, 0.0,0.0,0.0, 0.0,1.0,1.0, 0.0,1.0, 1.0,1.0,1.0, 0.0]=>0(expected0)

[1.0,1.0, 1.0,1.0,0.0,0.0,0.0,1.0,1.0, 1.0,1.0, 1.0,1.0, 1.0,1.0, 0.0,1.0, 1.0,0.0, 0.0,0.0, 1.0, 0.0,0.0,0.0, 0.0,1.0,0.0, 0.0,1.0, 1.0,1.0,0.0, 1.0]=>1(expected1)

[1.0,0.0, 1.0,1.0,0.0,0.0,0.0,1.0,1.0, 1.0,1.0, 0.0,1.0, 1.0,0.0, 0.0,1.0, 1.0,0.0, 1.0,0.0, 1.0, 0.0,0.0,0.0, 0.0,1.0,1.0, 0.0,1.0, 1.0,1.0,1.0, 0.0]=>0(expected1)

[1.0,1.0, 1.0,1.0,0.0,0.0,0.0,1.0, 1.0,1.0, 1.0,0.0, 1.0, 1.0,1.0, 0.0,1.0, 1.0,0.0, 1.0,0.0, 1.0, 0.0,0.0,0.0, 0.0,1.0,1.0, 0.0,1.0, 1.0,1.0,1.0, 0.0]=>0(expected0)

[1.0, 0.0, 1.0,1.0,0.0,0.0,0.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 0.0, 1.0, 0.0, 0.0, 1.0, 0.0, 1.0, 0.0,0.0,0.0, 0.0,1.0,1.0, 0.0,1.0, 1.0,1.0,1.0, 0.0]=>0(expected0)

[1.0,1.0, 1.0,1.0,1.0,0.0,0.0,1.0, 1.0, 1.0,1.0, 0.0,1.0, 1.0,1.0, 0.0,1.0, 0.0, 0.0,0.0, 0.0, 1.0,0.0, 0.0,1.0,1.0,1.0,0.0,0.0,1.0, 1.0, 1.0, 1.0, 0.0]=>0(expected0)

[1.0,0.0, 1.0,1.0,0.0,0.0,0.0,1.0,1.0, 1.0,1.0, 0.0,1.0, 1.0,1.0, 0.0,1.0, 0.0,0.0, 1.0,0.0, 1.0, 0.0,0.0,0.0, 0.0,1.0,1.0, 0.0,1.0, 1.0,1.0,1.0, 0.0]=>0(expected0)

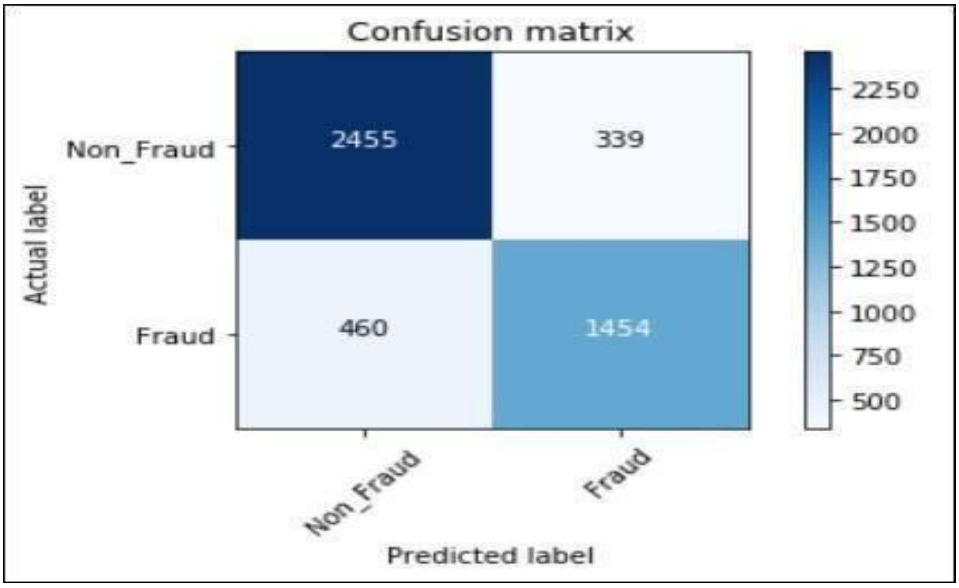
[1.0,0.0, 1.0,1.0,1.0,0.0,0.0,1.0,1.0, 1.0,1.0, 1.0,1.0, 1.0,1.0, 0.0,1.0, 0.0,0.0, 1.0,0.0, 1.0, 0.0,0.0,0.0, 1.0,1.0,0.0, 0.0,1.0, 1.0,1.0,1.0, 0.0]=>1(expected0)

[1.0,0.0, 1.0,1.0,0.0,0.0,0.0,1.0, 1.0,1.0, 1.0,0.0, 1.0,1.0,1.0, 0.0,1.0, 1.0,0.0, 1.0,0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 1.0, 1.0, 0.0, 1.0, 0.0, 1.0, 1.0, 0.0]=>0(expected0)[1.0,0.0,1.0,1.0, 0.0, 0.0, 0.0,1.0,1.0,1.0,1.0,0.0,1.0,1.0,0.0,0.0,1.0,1.0,0.0,1.0,0.0,1.0,0.0,0.0,0.0,0.0,1.0,1.0, 0.0,1.0,1.0, 1.0,1.0, 0.0]=>0(expected1)

[1.0,1.0, 1.0,1.0,0.0,0.0,0.0,1.0,1.0, 1.0,1.0, 0.0,1.0, 1.0,1.0, 0.0,1.0, 0.0,1.0, 1.0,0.0, 1.0, 0.0,0.0,0.0, 0.0,1.0,1.0, 0.0,1.0, 1.0,1.0,1.0, 0.0]=>0(expected0)

[1.0,0.0, 1.0,1.0,0.0,0.0,0.0,1.0, 1.0,1.0, 1.0,0.0, 1.0, 1.0,1.0, 0.0,1.0, 1.0,0.0, 1.0,0.0, 1.0, 0.0,0.0,0.0,0.0,1.0,1.0,0.0,1.0,1.0,1.0,1.0,0.0]=>0(expected0)

AppendixE: ConfusionMatrixandClassificationReport



```
classification_report:
              precision    recall  f1-score   support

       fraud       0.00      0.00      0.00        41
    nonfraud       0.99      1.00      1.00      4667

   micro avg       0.99      0.99      0.99      4708
   macro avg       0.50      0.50      0.50      4708
weighted avg       0.98      0.99      0.99      4708
 samples avg       0.99      0.99      0.99      4708
```

AppendixF:ModelSummary

```
          Model summary about CNN Algorithm:

Model: "sequential_12"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_22 (Conv2D)           (None, 6, 6, 64)          128
_________________________________________________________________
conv2d_23 (Conv2D)           (None, 6, 6, 64)          4160
_________________________________________________________________
max_pooling2d_14 (MaxPooling (None, 6, 6, 64)          0
_________________________________________________________________
dropout_15 (Dropout)         (None, 6, 6, 64)          0
_________________________________________________________________
flatten_11 (Flatten)         (None, 2304)              0
_________________________________________________________________
dense_20 (Dense)             (None, 128)               295040
_________________________________________________________________
dropout_16 (Dropout)         (None, 128)               0
_________________________________________________________________
dense_21 (Dense)             (None, 2)                 258
=================================================================
Total params: 299,586
Trainable params: 299,586
Non-trainable params: 0
```

AppendixG:Interview

1. Howdotheauditorsauditthetaxpayers?
2. Whichtaxpayersgetpriorityfromtheauditors?
3. Howistheprocessoftheauditingtask?
4. Whatisthemaintoolusedbyauditorsduringauditactivities?
5. Whatisthecurrentactivitytoprotectsuspectsoffraud?
6. Whatarethecriteriatodetect fraudstersinyourorganization?