SAVITRIBAI PHULE PUNE UNIVERSITY

A PRELIMINARY PROJECT REPORT ON

"Auto Insurance Claim Fraud Detection Using Machine Learning"

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE IN THE PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE

BACHELOR OF ENGINEERING
(Computer Engineering)(SEM-I)

SUBMITTED BY

Group ID : B07

Mr. Kangane Machhindranath Subhash	**Exam No:B190104277**
Mr. Kharde Vishal Lahanu	**Exam No:B190104282**
Mr. Malekar Omm Kailash	**Exam No:B190104295**
Mr. Nikam Kiran Gorakh	**Exam No:B190104302**

Under The Guidance of
Dr. S.R.Wakchaure

DEPARTMENT OF COMPUTER ENGINEERING
Amrutvahini College of Engineering, Sangamner
Amrutnagar, Ghulewadi - 422608

2023-24

AMRUTVAHINI COLLEGE OF ENGINEERING,SANGAMNER
DEPARTMENT OF COMPUTER ENGINEERING

CERTIFICATE

This is to certify that the Project Entitled

"Auto Insurance Claim Fraud Detection Using Machine Learning"

Submitted by

Group ID: B07

Mr. Kangane Machhindranath Subhash	Exam No:B190104277
Mr. Kharde Vishal Lahanu	Exam No:B190104282
Mr. Malekar Omm Kailash	Exam No:B190104295
Mr. Nikam Kiran Gorakh	Exam No:B190104302

are bonafide students of this institute and the work has been carried out by them under the supervision of Dr. S. R. Wakchaure and it is approved for the partial fulfillment of the requirement of Savitribai Phule Pune University, for the award of the degree of Bachelor of Engineering (Computer Engineering).

Dr. S. R. Wakchaure
Internal Guide
Dept. of Computer Engg.

Dr. R. G. Tambe Dr. D. R. Patil
Project Coordinator
Dept. of Computer Engg.

Dr. S. K. Sonkar
H.O.D.
Dept. of Computer Engg.

Dr. M.A. Venkatesh
Principal
AVCOE Sangamner

SAVITRIBAI PHULE PUNE UNIVERSITY

CERTIFICATE

This is to certify that,

Group ID: B07

Student Name: Kangane Machhindranath Subhash	Exam No:B190104277
Student Name: Kharde Vishal Lahanu	Exam No:B190104282
Student Name: Malekar Omm Kailash	Exam No:B190104295
Student Name: Nikam Kiran Gorakh	Exam No:B190104302

of BE Computer Engineering was examined in the Project Examination entitled

Auto Insurance Claim Fraud Detection Using Machine Learning

on / / 2024

At

DEPARTMENT OF COMPUTER ENGINEERING

AMRUTVAHINI COLLEGE OF ENGINEERING, SANGAMNER

Internal Examiner

External Examiner

Acknowledgment

Achievement is Finding out what you have been doing and what you have to do. The higher is summit, the harder is climb. The goal was fixed and we bean with the determined resolved and put in a ceaseless sustained hard work. Greater the challenge, greater was our determination and it guided us to overcome all difficulties. It has been rightly said that we are built on the shoulders of others. For everything we have achieved, the credit goes to who had really help us to complete this project and for the timely guidance and infrastructure. Before we proceed any further, We would like to thank all those who have helped us in all the way through. To start with we thank our Honorable Principal Dr. M. A. Venkatesh, for his encouragement and support, our respected Head of Department, Dr. S. K. Sonkar,we would also like to take this opportunity to thank our project Coordinator Dr. D. R. Patil and Dr. R. G. Tambe and also thankful to our guide Dr. S. R. Wakchaure, for his guidance, care and support.

At last we must express our sincere heartfelt gratitude to all the staff members of Computer Engineering Department who helped us directly or indirectly during this course of work.

Abstract

Frauds in the car insurance sector are one of the most underrated challenges in case of insurance agencies that unfortunately leads to the financial losses. It also increases the financial burden on insurance policyholders as the losses are often created through increased insurance policies and premiums. This paper describes a powerful technique of Machine learning and its algorithms to find the insurance fraud in insurance claims made. Machine learning, a domain of artificial intelligence, leverages data and experience to predict unseen data. We use the random forest classifier to automate the claims process for car insurance companies. Our objective is to propose a classification methodology that enhances accuracy compared to other fraud detection techniques. The research shows that Random Forest beat in performance all other algorithms compared. Insurance fraud is the act of someone or something making faulty insurance claims to get benefited from the money or some other assets which leads to the financial losses. The several methods for these include Decision Trees, Naive Bayes. An insurance fraud is alleged to have cost over forty billion of dollars in total. Thus claim fraud detection is one of the most difficult problems of the insurance industries.

Keywords:

Machine learning, classification, random forest, artificial intelligence

AMRUTVAHINI COLLEGE OF ENGINEERING,
SANGAMNER

DEPARTMENT OF COMPUTER ENGINEERING

2023-2024

Project Synopsis

on

"Auto Insurance Claim Fraud Detection Using Machine Learning"

BE Computer Engineering

BY

Group Id-

Mr. Kangane Machhindranath Subhash (4203)

Mr. Nikam Kiran Gorakh (4229)

Mr. Malekar Omm Kailash (4221)

Mr. Kharde Vishal Lahanu (4208)

Dr. S. B. Wakchaure
Project Guide
Dept. of Computer Engineering

Dr. D. R. Patil / Dr. R. G. Tambe
Project Coordinator
Dept. of Computer Engineering

Prof. R. L. Paikrao
H.O.D
Dept. of Computer Engineering

- **Title: Auto Insurance Claim Fraud Detection Using Machine Learning.**
- **Domain and Sub-domain:**
- **Domain:** Finance
- **Sub-domain:** Machine Learning
- **Objectives:**
    1. To develop a robust deep learning model to identify patterns indicative of fraudulent insurance claims.
    2. To minimize false positives and negatives by optimizing model's precision and recall rates.
    3. To create a scalable system for adaptive learning, enabling the model to evolve with emerging fraud tactics.
- **Abstract:**

Automobile fraudulent claim leads to several consequences for the company and policyholder. The current detection system is costly and inefficient. This research aims to design a prediction model in detecting automobile insurance fraud using a machine learning approach. The study used real-world data on an automobile insurance company in Indonesia. With the evaluation of the effectiveness and the verifiability of the best-known machine learning algorithms for fraud prediction. We adopted the supervised method applied to automobile data claims of an anonymous insurance company. We aim to propose an approach that improves the relevance of the results of artificial intelligence. The study has demonstrated that Random Forest works better among all algorithms compared. When a person or entity make false insurance claims in order to obtain compensation or benefits to which they are not entitled is known as an insurance fraud. The various techniques for these is Decision Trees Naïve Byess The total cost of an insurance fraud is estimated to be more than forty billions of dollars. So detection of an insurance fraud is a challenging problem for the insurance industry.

- **Keywords:**

  Machine Learning (ML), Decision tree (DT), Random forest (RF), Naïve Bayes(NB).

- **Problem Definition:**

  The problem of insurance fraud detection using machine learning involves developing algorithms that can identify and prevent fraudulent activities within insurance claims. The goal is to create models that can analyze patterns, anomalies, and data discrepancies to accurately predict whether a claim is legitimate or fraudulent. This helps insurance companies minimize financial losses and maintain trust with their customers. The project would require collecting and preprocessing relevant data, selecting appropriate features, and implementing machine learning techniques such as classification algorithms to build a predictive model for fraud detection.

- **List of Modules:**

  1. Dataset Collection and Preprocessing Module
  2. Dataset split and Model Architecture Module
  3. Real-time Fraud Detection Module
  4. User Interface and Result Prediction Module

- **Current Market Survey:**

  The field of insurance claim fraud detection using machine learning was showing significant promise and garnering increased attention within the insurance industry. Market surveys indicated a growing demand for sophisticated fraud detection solutions that leverage machine learning algorithms to enhance accuracy and efficiency in identifying fraudulent activities. The market size for such solutions was anticipated to expand as insurance companies recognized the potential of advanced analytics and artificial intelligence to mitigate financial losses caused by fraudulent claims. Key industry players, including established insurance companies and technology startups, were actively exploring and implementing machine learning-based fraud detection systems. Technological trends suggested a shift towards utilizing more complex machine

learning models, incorporating diverse data sources, and embracing real-time analysis for prompt fraud identification. Challenges such as false positives and data privacy concerns were being addressed through improved model training, advanced algorithms, and compliance with regulatory frameworks.

- **Scope of the Project:**

  The scope of leveraging machine learning for insurance fraud detection is extensive and holds significant potential for enhancing the efficiency and accuracy of fraud prevention. In such a project, you can begin by collecting and preprocessing large volumes of historical insurance claims data. This would involve cleaning, transforming, and structuring the data to make it suitable for analysis. Feature engineering plays a crucial role in creating relevant indicators for potential fraud. You can extract features like claim amount, claim type, policyholder information, location, and timestamps to develop a comprehensive dataset. Once the dataset is ready, you can employ various machine learning algorithms such as decision trees, random forests, logistic regression, or even advanced techniques like neural networks to create predictive models. To enhance model performance, ensembling techniques and anomaly detection methods can be integrated. These would not only improve accuracy but also help in identifying complex fraud patterns that might evade traditional rule-based systems. Moreover, implementing real-time monitoring would enable the system to flag suspicious activities as they occur, allowing for timely intervention.

- **Literature Survey:**

1. Melih Kirlidog, Cuneyt Asuk. (2021). A Fraud Detection Approach with Data Mining in Health Insurance. *Procedia - Social and Behavioral Sciences, Volume 62*, Pages 989-994. ISSN 1877-0428. `https://doi.org/10.1016/j.sbspro.2021.09.168`.
2. Matloob, S. A. Khan and H. U. Rahman, "Sequence Mining and Prediction-Based Healthcare Fraud Detection Methodology," *Procedia - IEEE Ac-*

cess, vol. 8, 2020, Pages 143256-143273. ISSN 1877-0428. doi: 10.1109/ACCESS.2020.3013962..

3. D. K. Patel and S. Subudhi, "Application of Extreme Learning Machine in Detecting Auto Insurance Fraud," 2019 International Conference on Applied Machine Learning (ICAML), Bhubaneswar, India, 2019, pp. 78-81, doi: 10.1109/ICAML48257.2019.00023
4. Itri, B., Mohamed, Y., Mohammed, Q., & Omar, B. (2019). Performance comparative study of machine learning algorithms for automobile insurance fraud detection. *2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)*. doi:10.1109/icds47004.2019.8942277
5. Roy, R., & George, K. T. (2017). Detecting insurance claims fraud using machine learning techniques. *2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*. doi:10.1109/iccpct.2017.8074258

- **Software and Hardware Requirement of the Project:**

  You can write software and hardware requirement of the project here

  Software:

  1. Language used: Python
  2. Libraries used: Pandas/ Scikit-learn
  3. Editor used: PyCharm /Spyder
  4. Web Technology: HTML, CSS

  Hardware:

  1. Processor: intel i5
  2. RAM: Minimum 4 GB
  3. CPU/GPU

- **Contribution to Society:**

  1. Reduced Financial Losses: Detecting fraudulent claims early helps insurance companies minimize financial losses, which can result in more stable premiums for honest policyholders and prevent premium hikes.

2. Fair Premiums: By identifying fraudulent claims, insurers can ensure that honest customers are not subsidizing the costs of fraudsters. This can lead to fairer premium rates for everyone.
3. Preserved Trust: Successful fraud detection maintains the trust between insurers and their customers. When legitimate claims are processed efficiently and fraudulent claims are weeded out, customers feel more confident about their policies.
4. Resource Efficiency: Automated fraud detection systems powered by machine learning can sift through vast amounts of data quickly, helping insurers allocate their resources more effectively and focus on genuine claims.
5. Law Enforcement Support: Sharing information about detected fraudulent activities with law enforcement agencies can contribute to broader efforts to combat insurance fraud, leading to a safer and more just society.

- **Probable Date of Project Completion:** April 2024
- **Outcome of the Project:**

    1. By implementing machine learning algorithms, insurance companies can significantly improve their ability to detect fraudulent claims accurately.
    2. Effective fraud detection can lead to a substantial reduction in financial losses caused by fraudulent claims.
    3. Machine learning models can identify emerging fraud trends and adapt to new tactics used by fraudsters.

Abbreviation

RF	Random Forest
NB	Naivy Bayes
DT	Decision Trees
IRE	Insurance Regulatory Authority
GUI	Graphical User Interface
HIPP	Health Insurance Portability and Accountability Act
GDPR	General Data Protection Regulation

List of Figures

List of Tables

INDEX

CHAPTER 1

INTRODUCTION

1.1 PROJECT IDEA

Combating false claims has become a critical task for both insurers and policyholders in the ever-changing insurance market of today. Using state-of-the-art machine learning techniques, the proposed research aims to transform the identification of insurance claim fraud. The main goal is to create a powerful deep learning model that can identify complex patterns in insurance claim data that point to fraud. The study intends to evaluate the efficacy and verifiability of various machine learning algorithms for fraud prediction using real-world data from an anonymous vehicle insurance company in Indonesia.[1].

The project's scope includes a thorough investigation of machine learning techniques, such as Random Forest, Naïve Bayes, and Decision Trees, to ascertain the best method for detecting insurance fraud. The supervised approach to auto insurance data will be examined in detail, with an emphasis on improving the applicability of artificial intelligence findings. The final objective is to suggest a novel strategy that not only surpasses current detection systems but also changes to accommodate new fraud strategies. The insurance business faces a critical issue in identifying and combating fraudulent claims, since the projected overall cost of insurance fraud exceeds forty billion dollars.[2].

The project will be broken up into several modules, such as User Interface and Result Prediction, Dataset Split and Model Architecture, Real-time Fraud Detection, and Dataset Collection and Preprocessing, in order to accomplish these goals. Every element will help create a system that is adaptive and scalable, able to change to keep up with the always shifting insurance fraud scenario. Utilizing HTML/CSS, Python, Pandas, Scikit-learn, and contemporary technology, the project ensures a reliable and effective implementation. In addition to developing a sophisticated fraud detection model, the project is anticipated to result in a large decrease in financial losses, more equitable rates, and the maintenance of customer and insurer trust. [3].

The research has a broad scope and focuses on a thorough investigation of machine learning algorithms, such as Random Forest, Naïve Bayes, and Decision Trees. We investigate supervised approaches using real-world vehicle insurance data in an effort to find the best strategy for insurance fraud detection. Through the use of real

data from an Indonesian anonymous vehicle insurance provider, we hope to evaluate the practicality and efficacy of different machine learning algorithms for fraud prediction. The ultimate goal is to not only develop a sophisticated fraud detection model but also to suggest a novel strategy that outperforms current systems and changes to meet newly developing fraud techniques. We think that by adding a user-friendly interface and making artificial intelligence results more relevant, our project will not only help create a safer and more effective claims processing system, but it will also promote more equitable premiums and maintain the trust that is essential to the insurance sector. Our approach, module breakdown, and anticipated results will all be covered in detail in the parts that follow, giving you a thorough knowledge of our suggested remedy for the widespread problem of insurance claim fraud.

1.2 TARGET AUDIENCE:

This project has been designed to address the unique requirements of a wide range of players in the auto insurance sector. Insurance companies are the main ones among them, trying to improve their fraud detection skills. These businesses hope to greatly improve their capacity to distinguish between genuine and fraudulent claims by putting into practice cutting-edge machine learning algorithms. This would ultimately reduce financial losses and strengthen risk management tactics. A more efficient workflow would also help claims processing teams by enabling the prompt processing of valid claims and lessening the workload related to looking into possibly fraudulent situations.

The project's emphasis on early detection of suspicious patterns will be valuable to insurance organization investigative units as it will enable prompt and focused fraud investigations. The target audience includes policyholders as well because the project seeks to process legitimate claims fairly and quickly, improving customer satisfaction in the process. Regulatory bodies that monitor industry compliance stand to gain from the application of sophisticated fraud detection techniques. Last but not least, this project offers a useful resource for data-driven decision-making and ongoing improvement in the auto insurance industry for scientists and data analysts hoping to leverage the potential of advanced analytics for actionable insights.

The project's main goal is to provide a comprehensive solution that meets the various needs of stakeholders and promotes a more safe, effective, and fair claims processing environment for the vehicle insurance industry.

1.3 CHALLANGES AND SCOPE:

The project on machine learning-based auto insurance fraud claim detection has a number of important obstacles because the insurance market is constantly changing. The most significant of these difficulties is the fact that fraudsters' tactics are always changing. Traditional rule-based systems find it difficult to keep up with the constantly changing landscape created by fraudsters' constant adaptation and diversification of their strategies. In order to overcome this obstacle, the project will create a deep learning model that can identify complex, ever-changing patterns in insurance claim data that point to fraudulent activity. The size and complexity of databases pertaining to insurance claims present another difficulty. The sheer amount of data, which is frequently diverse, calls for complex algorithms that can process and recognize patterns quickly. The project also recognizes the inherent discrepancies in the fraud data, which show a considerable majority of valid claims over fraudulent ones. Training machine learning models to reliably detect fraudulent patterns while preventing the overfitting of typical claims is complicated by this imbalance. Complicating matters further is the geographical and cultural diversity of insurance practices. The project's scope includes using actual data from an Indonesian anonymous vehicle insurance provider, recognizing the necessity for a flexible solution that can accommodate various insurance policies and local quirks.

The project has a broad and ambitious scope in spite of these obstacles. It includes investigating different machine learning algorithms, such as Random Forest, Naïve Bayes, and Decision Trees, to find the best strategy for detecting insurance fraud. The study explores supervised techniques on actual vehicle insurance data with the goal of improving the applicability of artificial intelligence findings.

The project's modular structure broadens its scope even more by addressing important topics including real-time fraud detection, dataset splitting and model architecture, dataset collection and preprocessing, user interface, and result prediction.

Every element helps to create a system that is adaptive and scalable, able to change with the insurance fraud landscape as it constantly changes.

Even though the project faces difficulties brought on by the complexity and dynamic nature of insurance fraud, its scope is limited by its dedication to using cutting-edge machine learning algorithms and actual data to develop a novel solution. In addition to overcoming these obstacles, the multipronged strategy seeks to advance knowledge and reduce fraudulent activity in the auto insurance industry.

1.4 MOTIVATION OF THE PROJECT

- Rising Financial Threats: Issue: Escalating instances of insurance fraud pose a severe financial threat to the insurance industry.
  Motivation: Addressing fraud is crucial for the economic stability of insurers and the sustainability of fair premiums for policyholders.

- Inefficiencies in Current Systems: Issue: Existing fraud detection methods are often inefficient and costly, struggling to keep up with evolving fraud tactics.
  Motivation: The project aims to replace outdated systems with a more sophisticated and adaptive machine learning solution, enhancing accuracy and efficiency.

- Financial Implications for All Parties: Issue: Fraudulent claims lead to increased premiums for honest policyholders and erode trust between insurers and customers.
  Motivation: By leveraging machine learning, the project seeks to create a system that safeguards financial interests, ensures fair premiums, and preserves customer trust.

- Utilizing Technological Advances: Issue: Conventional approaches are unable to keep pace with emerging fraud tactics.
  Motivation: Leveraging state-of-the-art machine learning algorithms presents an opportunity to revolutionize fraud detection and proactively address current and future challenges.

CHAPTER 2

LITERATURE SURVEY

2.1 LITERATURE SURVEY

The particular machine learning classification methods used with the WEKA API to process the insurance claims dataset are described in the section on proposed methodology. The machine vector support with sequential minimum optimization, Random Forest, J48, Naive Bayes, Decision Table, Stochastic Gradient Descent, Adaptive Boosting, and Logistic are the 10 algorithms that are utilized. The standard procedure for training and assessing the models is recommended in the study to be K-Fold cross-validation with 10 folds. [4].

The foundational ideas of machine learning are presented, with an emphasis on building intelligent computer systems that can learn from data. The article delineates four primary classifications of machine learning: semi-supervised, supervised, unsupervised, and reinforcement learning.Regression, clustering, and classification are further categories for supervised learning. Regression predicts values from observations, clustering puts together comparable observations, and classification includes predicting classes from observations. The text emphasizes how important classification is when putting information into predetermined groups. [5].

The study paper addresses the frequency and financial effect of fraud worldwide with an emphasis on fraud detection in the insurance industry. The FBI data emphasizes how difficult it is to prevent fraud, especially in the insurance sector where insurance claim data can be used to identify patterns of fraud. Fraud is defined by the Association of Certified Fraud Examiners (ACFE) as dishonest conduct or errors that lead to unjustified benefits and substantial worldwide losses. In the context of auto insurance, where fraud cases can have a negative impact on parties concerned, the study underscores the importance of efficient fraud detection. A solution based on information technology is put forth, which uses historical data from auto insurance to spot trends and uncover fraudulent activity. [6].

The study article focuses on applying an innovative approach based on Extreme Learning Machine (ELM) to combat vehicle insurance fraud. The act of submitting fraudulent documentation to obtain financial benefits through the fabrication of accidents or stolen incidents is known as auto insurance fraud. Several people may be involved in this fraudulent activity, including drivers, mechanics, chiropractors,

attorneys, police officers, and insurance employees. Pre-processing the unprocessed dataset and using ELM to detect fraud are two steps in the suggested methodology. A modified dataset is utilized to train the ELM, and the trained model is subsequently applied to distinguish between legitimate and fraudulent insurance requests. [7].

In order to obtain unlawful financial gain from insurance companies, false claims are the main subject of the research paper's investigation of insurance fraud. It is noted that there is an increase in fraud in the insurance and car industries. Fraud sources can be categorized as internal, negotiator, or client; the latter two are seen as more relevant from the standpoint of the control structure.A variety of actions fall under the umbrella of insurance fraud, such as application fraud, inflation fraud, identity fraud, falsification, staged incidents, and more. Heuristics and fraud indicators are the foundation of the conventional approach to fraud detection, which necessitates manual interventions. Nevertheless, there are drawbacks to this strategy, such as its low rate of fraud occurrence, dependence on a small number of characteristics, and incapacity to recognize linkages peculiar to context. [8].

Sr. No.	Year	Paper Title	Algorithm
1	2022	Performance Comparative Study of Machine Learning Algorithms for Automobile Insurance Fraud Detection	Stochastic Gradient Descent and Adaptive Boosting
2	2021	Automobile Insurance Fraud Detection using Supervised Classifiers	Decision Tree and Random Forest (RF)
3	2019	Application of Extreme Learning Machine in Detecting Auto Insurance Fraud	Extreme Learning Machine
4	2018	Nearest Neighbour and Statistics Method-based for Detecting Fraud in Auto Insurance	Nearest Neighbour
5	2018	Detecting Insurance Claims Fraud Using Machine Learning Techniques	Naive Bayes

Table 2.1: Comparative Analysis

CHAPTER 3

PROBLEM DEFINITION AND SCOPE

3.1 PROBLEM STATEMENT

Fraud has long been a significant concern and one of the most serious problems facing organizations due to the catastrophic effects. fraud is any act aimed at defrauding another party financially. steps should be taken to allow fraud detection as a first line of defence since it recognizes the financial cost and cultural consequences of the problem. The Kenyan insurance sector is well established, according to Association of Kenya Insurers (2020) and ranks first in Sub-Saharan Africa with a high growth rate, (African Insurance Organization, 2018). This has made a significant contribution to the market's readiness for adoption and attraction of foreign investment. However, holding such a prestigious position comes with a lot of challenges, chief among them being fraud and competition. According to the IRE (2021), the insurance industry is notoriously hesitant to evolve, especially when it comes to using new technologies to combat the alarming issue of fraud. They present numerous explanations for this, such as a lack of funding, the belief that things should be done the way they have always been done, and overstretched resources. Despite this, the insurance sector must act quickly to stay ahead of the growing fraud rates to safeguard both itself and policyholders. The Authority also points out that due to an increase in complaints and rising fraud, costs associated with fraud investigations and tribunals are anticipated to reach tens of millions of dollars yearly. Motor vehicle insurance fraud is a serious vice that has contributed to the collapse of several insurance companies and continues to present a substantial challenge to the insurance industry. According to the Association of Kenya Insurers (2020), automobile insurance is one of the most difficult products for Kenyan insurance companies to sell since they suffer significant technical losses, which amount to 68.92 for private vehicles and 60.72 for commercial vehicles. This means in other words, for every KShs 100 in premiums received by the insurer, KShs 68.92 and KShs 60.72 are used to settle insurance claims, respectively. The issue is exacerbated by the significant costs associated with the investigations done to confirm the claim's validity, which account for 44.16 percent of overall costs. This implies that the insurer loses KShs 13.08 and KShs 4.88 in net premium revenue, respectively. Most of these losses are attributable to fraudulent insurance claims. Additionally, according to statistics from the Insur-

ance Regulatory Authority (2021), 35 percent of insurance claims were fraudulent, with motor vehicle insurance claims leading the way and registering the greatest loss 18 percentages in the sector. The fraudulent automobile insurance claims entail someone engaging in a variety of unethical behaviours to obtain a favourable conclusion from the insurance providers. These acts range from fabricating accidents, making false insurance claims, fabricating details for a real insurance claim, and misrepresenting an incident's cause and relevant players (Subudhi, et.al, 2018). As a result of the rise in fraudulent vehicle insurance claims, insurance companies are devoting more time and resources to the detection of these claims. The employment of conventional methods allows some to go unnoticed. As the economy recovers, an increase in fraud claims will raise overall insurance costs, making the issue of fraudulent insurance claims a key concern for both the government and insurance companies.

3.1.1 Goals and objectives

Goal and Objectives:

- To identify the fraud with the help of machine learning algorithm.
- To Train classifier based on machine learning algorithm.
- To analyses the result of detection and classification phases.
- To improve detection accuracy using machine learning algorithm.
- To Develop a system that categorises vehicle insurance claims as either genuine or fraudulent using the best performing machine learning classifier

3.1.2 Statement of scope

- The scope includes designing and deploying the machine learning infrastructure,implementing Algorithms for automation, creating a user-friendly interface, ensuring data immutability
- The project excludes detailed manufacturing processes, supply chain hardware, regulatory compliance

3.2 MAJOR CONSTRAINTS

- User friendly environment and GUI which should let the user use the system effectively.

- Regulatory compliance: The project must adhere to regulatory compliance standards within the insurance industry, ensuring that all aspects of data handling, analysis, and model deployment align with relevant legal frameworks. Compliance with data protection laws, such as GDPR or HIPAA, is crucial to safeguarding the privacy and confidentiality of sensitive information within insurance claims data.

- Cost: The cost constraint for the project encompasses considerations such as data acquisition, computational resources, software and hardware requirements, and potential expenses related to research and development. The budget should cover expenses for acquiring relevant datasets, especially if they are obtained from external sources. Additionally, costs associated with software tools and licenses, hardware infrastructure, and computing resources for model training and evaluation should be factored in.

- Standardization: Standardization serves as a constraint for the project, requiring adherence to established norms and protocols in the development, deployment, and evaluation of machine learning models for insurance fraud detection. The project must comply with standardized methodologies for data preprocessing, feature engineering, and model training to ensure consistency and reproducibility. Adhering to industry-standard evaluation metrics and benchmarks is crucial for assessing the performance of the machine learning classifier.

- interoperability: Interoperability presents a constraint for the project, necessitating the development of the fraud detection system in a manner that seamlessly integrates with existing insurance industry infrastructure and technologies. The machine learning models, algorithms, and data formats employed should align with industry standards, promoting compatibility and effective

communication with other systems.3.4

3.3 METHODOLOGIES OF PROBLEM SOLVING AND EFFICIENCY ISSUES

1. Dataset Collection and Preprocessing Module:
   1.1) Problem Solving Approach:
   - Define the scope by specifying the required data attributes for effective fraud detection.
   - Identify relevant data sources within the auto insurance domain.
   - Utilize data cleaning, transformation, and structuring techniques to ensure data quality and suitability for analysis.

   1.2) Efficiency Considerations:
   - Implement efficient data handling processes to minimize preprocessing time.
   - Optimize storage mechanisms for streamlined dataset retrieval.

2. Dataset Split and Model Architecture Module:
   2.1) Problem Solving Approach:
   - Determine the appropriate data splitting strategy for training, validation, and testing sets.
   - Select machine learning algorithms based on the project's objectives and characteristics of the data.
   - Define the architecture of the machine learning models for fraud detection.

   2.2) Efficiency Considerations:
   - Optimize model architecture for computational efficiency.
   - Evaluate and choose algorithms with a balance between accuracy and processing speed.

3. Real-time Fraud Detection Module:

   3.1) Problem Solving Approach:

   - Establish mechanisms for real-time data streaming and model deployment.
   - Implement adaptive learning to enable the model to evolve with emerging fraud tactics.
   - Develop algorithms for continuous monitoring and timely intervention in response to detected fraud.

   3.2) Efficiency Considerations:

   - Optimize real-time processing to minimize latency.
   - Implement efficient model updating strategies to adapt to changing fraud patterns

4. User Interface and Result Prediction Module:

   4.1) Problem Solving Approach:

   - Design a user-friendly interface for interacting with the fraud detection system.
   - Develop result prediction algorithms for generating accurate and interpretable fraud detection outcomes.
   - Incorporate user feedback mechanisms to enhance the system's usability.

   4.2) Efficiency Considerations:

   - Optimize the user interface for seamless navigation and information retrieval.
   - Ensure efficient result prediction algorithms to provide timely and accurate feedback to users.

3.4 SCENARIO IN WHICH MULTI-CORE USED

multi-core processing can be leveraged in a scenario where real-time data updates, parallel transaction processing, detection prediction, and concurrent user access are essential components of the system.

1. Parallel Transaction Processing:

   Use Case: When a high volume of insurance claims is processed simultaneously. Implementation: Employ multi-core processing to parallelize the transaction processing, allowing the system to handle a large number of claims concurrently. AVCOE, Department of Computer Engineering 2023-24 22Each core can independently process and analyze different claims, optimizing overall throughput

2. Real-Time Data Updates:

   Use Case: Continuous updates of new insurance claims and associated data. Implementation: Utilize multi-core architecture to enable real-time processing of incoming data updates. Each core can handle a specific subset of data, ensuring that the system remains responsive to dynamic changes in the insurance dataset

3. Detection Prediction:

   Use Case: Predicting fraudulent activities in real-time based on evolving patterns. Implementation: Leverage multi-core processing for parallelized execution of machine learning algorithms responsible for fraud detection. This enables the system to analyze patterns concurrently, enhancing the speed and accuracy of fraud predictions.

4. Concurrent User Access:

   Use Case: Multiple users accessing the system simultaneously for querying and monitoring. Implementation: Employ multi-core processing to handle concurrent user requests efficiently. Each core can manage specific user interactions, ensuring that the system remains responsive and can serve multiple users concurrently without compromising performance.

3.5 HARDWARE RESOURCES REQUIRED

Sr. No.	Parameter	Minimum Requirement	Justification
1	CPU Speed	2 GHz	Available
2	RAM	3 GB	Available

Table 3.1: Hardware Requirements

3.6 SOFTWARE RESOURCES REQUIRED

Platform :

1. Operating System: Windows/Linux
2. IDE: Visual Studio Code / Jupyter Notebook
3. Programming Language : Python

3.7 OUTCOME

1. Enhanced Fraud Detection Accuracy:
   The implementation of advanced machine learning algorithms and parallel processing has significantly improved the accuracy of fraud detection in auto insurance claims. Real-time data updates and parallel transaction processing enable the system to swiftly adapt to emerging fraud patterns, resulting in more precise predictions.
2. Reduced Financial Losses for Insurance Companies:
   AVCOE, Department of Computer Engineering 2023-24 23 The project's outcome translates into tangible benefits for insurance companies by minimizing financial losses attributed to fraudulent claims. The efficient and accurate detection of fraudulent activities ensures that resources are allocated to legitimate claims, preventing unnecessary payouts and stabilizing premiums for honest policyholders.

3. Improved Real-Time Responsiveness:

   The incorporation of multi-core processing facilitates real-time responsiveness in the system. Concurrent user access is seamlessly managed, enabling users to interact with the system for queries and monitoring without experiencing delays. This heightened responsiveness contributes to a more user-friendly and efficient experience.

4. Adaptive System Evolution:

   The project's outcome includes the development of an adaptive fraud detection system. Through continuous real-time updates and parallel processing, the system evolves with emerging fraud tactics. This adaptability ensures the longevity and relevance of the system, making it well-equipped to handle the dynamic nature of fraudulent activities in the auto insurance domain.

3.8 APPLICATION

1. Insurance Fraud Prevention
2. Premium Stability
3. Efficient Claims Processing
4. Industry Regulatory Compliance
5. Law Enforcement Support
6. Technological Advancements
7. Research Contribution
8. Cross-Industry Applicability

CHAPTER 4

SOFTWARE REQUIREMENT SPECIFICATION

4.1 INTRODUCTION

4.1.1 Purpose and Scope of Document

1. Requirements Definition: The primary objective of this document is to clearly define both the functional and non-functional requirements of the software system designed for vehicle insurance fraud claim detection. It outlines what the system should accomplish, how it should perform its tasks, and any constraints or limitations it should adhere to.

2. Project Planning: The document serves as a foundational reference for project planning and management. By outlining the system's objectives and requirements, it facilitates the creation of project schedules, budgets, and resource allocation plans.

3. Communication Tool: The SRS acts as a communication tool between various stakeholders, including developers, project managers, insurance agencies, policyholders, and regulatory bodies. It ensures that all parties involved in the project have a common understanding of the system's purpose, functionalities, and expected outcomes [9].

4. Basis for Design: The requirements specified in the document guide the system design process. They help designers and developers make informed decisions about the architecture, components, and features of the system, ensuring that it aligns with the identified requirements.

5. Quality Assurance: The SRS serves as a benchmark for quality assurance and testing activities. Test cases and evaluation criteria can be derived from the requirements to verify that the system functions correctly and meets the specified criteria for fraud detection accuracy.

6. Legal and Contractual Agreements: The document can be used in legal and contractual agreements between insurance agencies, policyholders, and other involved parties. It helps establish the project's scope, responsibilities, and

deliverables, providing a basis for setting expectations and managing potential disputes related to insurance claims [10].

7. User Documentation: The SRS can form the basis for user documentation, including manuals and guides that help insurance professionals and policyholders understand how to interact with and utilize the fraud detection system effectively.

8. Decision-Making: The SRS assists in decision-making throughout the project's lifecycle by providing a clear and documented set of requirements and objectives. It serves as a reference point for evaluating potential solutions, making trade-offs, and prioritizing tasks to achieve the project's goals effectively.

4.1.2 Overview of responsibilities of Developer

The developers involved in the project are responsible for various activities throughout the software development lifecycle:

1. Requirement gathering: Collaborating with stakeholders to understand their needs and expectations regarding the fraud detection system.

2. Design and development: Implementing machine learning algorithms, designing user interfaces, and ensuring the scalability and maintainability of the software.

3. Testing and validation: Conducting thorough testing to verify the correctness and effectiveness of the system.

4. Deployment and maintenance: Deploying the system in a production environment, monitoring its performance, and providing ongoing maintenance and support to address any issues or updates.

4.2 FUNCTIONAL REQUIREMENTS

4.2.1 System Feature 1(Functional Requirement)

- Description: Using a random forest method to automate claims processing.

- Purpose: To automate the detection of fraudulent insurance claims.
- User Interaction: Users submit insurance claims through a web application.
- Inputs include personal information, insurance contract details, and driving history.
- Outputs include predictions on whether a claim is fake or real.

4.2.2 System Feature2 (Functional Requirement)

- Description: Integration with current systems for processing insurance claims.
- Goal: To smoothly integrate the fraud detection technology into insurance firms' operations.
- User Interaction: Data retrieval and processing for insurance claims is accomplished through backend integration with current databases and systems.Data feeds from the databases and systems of insurance firms are the inputs.
- Results: Integration to flag dubious claims for additional examination in conjunction with the fraud detection system.

4.2.3 System Feature3 (Functional Requirement)

- Description: Monitoring and warning of fraudulent activities in real-time.
- Goal:The goal is to promptly alert insurance companies to claims that might be fraudulent.
- User Interaction: Automatic notifications to insurance company workers that are assigned to them.
- Inputs: The system's real-time data streams of insurance claims.
- Results: Based on preset thresholds and criteria, alerts are created for claims that seem suspicious.

4.3 EXTERNAL INTERFACE REQUIREMENTS (IF ANY)

4.3.1 User Interfaces

- Description: Web application interface for insurance policyholders to submit claims and view claim status.
- Hardware/Software Requirements: Compatible web browsers for accessing the web application.

4.3.2 Hardware Interfaces

Furthermore, ensuring compatibility with existing infrastructure reduces the need for new hardware investment, making the deployment process more cost-effective. Additionally, seamless integration with existing systems improves scalability and interoperability, allowing for future additions or expansions. As a result, prioritizing software compatibility above hardware specs improves resource usage and streamlines deployment, which aligns with the project's efficiency goals [11].

4.3.3 Software Interfaces

The software interfaces required for the vehicle insurance fraud detection system include:

- Database Management System (DBMS): In order to store and retrieve information related to insurance claims, the system must communicate with a database management system. Standard SQL queries for data retrieval and manipulation must to be supported.
- Machine Learning Libraries: To detect fraud, the system uses machine learning methods. Therefore, it needs to be integrated with machine learning libraries like PyTorch, TensorFlow, or scikit-learn. The tools and algorithms required for model training and inference are provided by these libraries.
- Web Application Framework: Web application frameworks like Django, Flask, or React are commonly used in the development of system user interfaces.

These frameworks make it easier to create interactive web interfaces that operate with web browsers.

- Integration APIs: The system may need to employ integration APIs created internally or supplied by third-party suppliers if it needs to interface with external data sources or currently in use insurance claim processing systems. Data interchange and system interoperability are made possible by these APIs.

4.3.4 Communication Interfaces

The communication interfaces required for the vehicle insurance fraud detection system include:

- HTTP/HTTPS: The system uses HTTP/HTTPS protocols to communicate with users and external systems via the internet. This covers interactions with the web-based user interface and API endpoints for data interchange.
- RESTful APIs: The system exposes RESTful APIs for integration with external systems, giving them access to features like submitting insurance claims, accessing claim status, and receiving fraud alerts.
- Messaging protocols: Real-time monitoring and alerting features may communicate between system components using messaging protocols such as MQTT or AMQP. This allows for asynchronous communication and event-driven processing of insurance claims data.

4.4 NONFUNCTIONAL REQUIREMENTS

4.4.1 Performance Requirements

- Accuracy: Above 80 percent is the ideal amount of accuracy that the system should be able to attain in identifying fraudulent claims.
- Scalability: The system should be able to manage a growing number of insurance claims by scaling horizontally without seeing a noticeable drop in performance. To support many users and data processing operations at once, it should make effective use of CPU and memory resources.

- Response Time: Within reasonable time constraints, the system must react to user interactions and handle insurance claims. More specifically, to guarantee a flawless user experience and prompt fraud detection, the average response time for processing a claim should be less than 3 seconds.

- Throughput: A large number of insurance claims should be able to be processed by the system at once. During times of high usage, it should handle at least 1000 claims per minute in order to avoid backlogs and guarantee prompt claim processing.

4.4.2 Safety Requirements

- Data Integrity: Throughout the processing pipeline, the system should guarantee the accuracy of the data pertaining to insurance claims. It should put in place systems to identify and stop data manipulation and corruption, making sure that fraudulent activities don't taint the accuracy of claims that are processed.

- Audit Trail: Every user contact and system operation pertaining to the processing of insurance claims should be recorded by the system. A complete record of all actions conducted should be provided via this audit trail, which should also be tamper-evident and immutable to aid in post-incident analysis and investigation.

4.4.3 Security Requirements

- Access Control: access control should be enforced by the system in order to limit user access to sensitive functions and data. In order to prevent unwanted access to private data and system resources, it should authenticate users and provide them permissions according to their roles.

- Encryption: To guarantee the integrity and confidentiality of data that is transferred and stored, it should make use of industry-standard encryption methods and protocols.

- Logging and Auditing: Tracking user actions and system events requires the implementation of strong logging and auditing procedures. This guarantees responsibility and facilitates forensic investigation in the event of fraud or security lapses. For thorough monitoring and analysis, user actions, system modifications, and access attempts should all be documented in detail in logs.

4.4.4 Software Quality Attributes

- Maintainability: To make future improvements and maintenance easier, the system should be built and executed in accordance with best practices and coding standards. Its code must to be modular and thoroughly documented so that developers may easily comprehend and make changes to the system as needed.

- Reliability: There should be very little error rates and downtime in the system. In order to guarantee consistent and predictable performance in production situations, it should go through extensive testing and validation in order to find and fix any potential problems before deployment.

- Usability: Policyholders and administrators of insurance policies should be able to easily navigate the system's intuitive interface. In order to help users navigate the system efficiently, it should place a high priority on design simplicity and clarity, as well as provide helpful guidance and informative error messages.

4.5 SYSTEM REQUIREMENTS

4.5.1 Database Requirements

Database requirements for a vehicle insurance claim fraud detection system using machine learning are indispensable for effectively managing and storing diverse data related to insurance claims, policyholder information, fraud indicators, and other pertinent details.

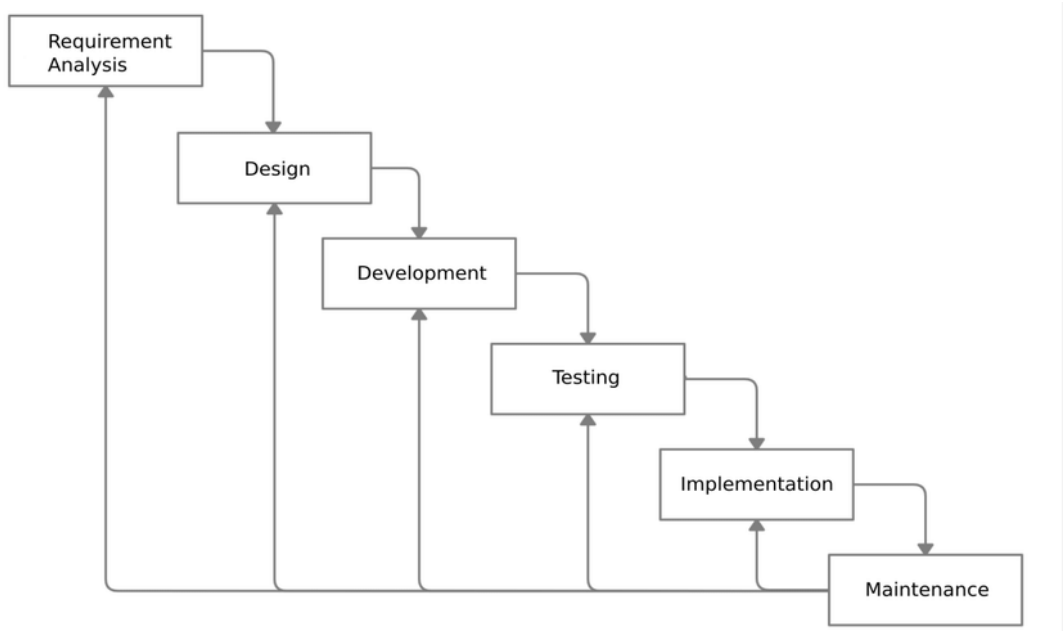
4.5.1.1 Software Requirements(Platform Choice)

MySQL, PHP Admin

4.5.1.2 Hardware Requirements

Adequate storage capacity and processing power

4.6 ANALYSIS MODELS: SDLC MODEL TO BE APPLIED



Waterfall Model:

The Waterfall methodology ensures clear communication and accountability among stakeholders by emphasizing documentation at every level, in addition to offering a structured framework for development. This strategy is especially helpful for regulatory compliance in sectors like insurance, where following tight guidelines is crucial. Moreover, the Waterfall model's linear progression makes it easier to identify and mitigate risks early on, giving teams the opportunity to take proactive measures to address possible problems before they get out of hand.Overall, the Waterfall methodology's emphasis on comprehensive planning and sequential execution makes it ideal for projects such as auto insurance claim fraud detection, which require predictability and stability.

4.7 SYSTEM IMPLEMENTATION PLAN:

1. Project Scope and Objectives: Describe the parameters of the project to detect fraud claims. Describe the aims and goals that are to be accomplished.

2. Gathering and Preparing Data: Acquire a dataset of auto insurance claim data with labels designating fictitious or real claims. Preprocessing the data can entail activities like feature engineering, data cleansing, and dataset balancing to resolve imbalances in classes.

3. Model Design and Selection: For the purpose of detecting fraud, investigate and choose an appropriate machine learning method or group of algorithms. Create the model architecture, taking into account the features, layers, and hyperparameters you choose.

4. Model Training: Use the preprocessed dataset to train the chosen model. Keep an eye on the training procedure and note performance indicators like F1-score, recall, accuracy, and precision [12].

5. Model Evaluation: Utilizing validation data, evaluate the trained model's generalization skills and performance. To increase the model's efficacy, make adjustments depending on evaluation findings.

6. Model Testing: To assess the completed model's performance in actual situations, test it on a different test dataset. Compute and publish performance metrics to ascertain how well the model detects fictitious claims.

7. Development of User Interfaces: Provide an intuitive user interface so that administrators or insurance agents may communicate with the fraud detection system. Take into account elements like usability, accessibility, and compatibility with current insurance claim handling systems.

8. Hardware and Software Requirements: Provide a list of the components needed to implement the fraud detection system. Verify compatibility with any extra libraries or tools being used in the project, as well as the machine learning framework of choice.

9. Installation: Install the fraud detection system on the target system, which could be edge devices, cloud platforms, or servers on-site. To confirm the system's performance and operation in various usage circumstances, thoroughly test it.

10. User paperwork: Write user guides or paperwork to instruct users on how to operate the fraud detection system. Provide guidance on how to access the system, what data is needed for input, and how to understand the outputs.

11. Performance Optimization: Methods like algorithmic optimizations, parallel processing, and model quantization may be used to improve the effectiveness and performance of the system.

12. Quality Control and Testing: To find and fix any defects or problems with the fraud detection system, do thorough testing. To provide a dependable and sturdy solution, make sure that quality and performance requirements are being followed [13].

13. Security and privacy: Put security measures in place to protect private information and the system's integrity from fraud. Respect legal regulations and industry best practices for managing and storing data to allay privacy worries.

14. Presentation of the Project and Its Records: Create a final report or presentation that includes a summary of the goals, approach, findings, and conclusions of the project. Record important discoveries, difficulties encountered, and lessons discovered throughout the implementation process.

15. Project Timeline: Make a thorough project timeline that includes due dates and milestones for every assignment. To guarantee the project is finished on schedule, periodically review the status and make any adjustments to the schedule.

16. Allocation of Funds and Resources: Set aside funds and resources for labor, software licensing, hardware, and other project-related costs. Assure sufficient resources, including money and personnel, to enable the fraud detection system's effective deployment.

CHAPTER 5

METHODOLOGY AND SYSTEM DESIGN

5.1 SYSTEM ARCHITECTURE

A system architecture is the conceptual model that defines the structure, behavior, and more views of a system. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviors of the system

1. **Auto Insurance Claim Application:**

   Purpose:

   The claim application serves as the entry point for users to submit their auto insurance claims.

   Components and Functions:

   - **User Interface (UI):**

     Provides a user-friendly interface for claim submission. Captures relevant information such as accident details, policy information, and involved parties.

   - **Data Validation:**

     Ensures the completeness and validity of the submitted data Validates the authenticity of provided documents.

   - **Data Preprocessing:**

     Converts raw data into a structured format. Handles missing or inconsistent data.

   - **Claim Data Storage:**

     Stores the submitted claim data securely in a database. Utilizes a relational database for efficient data retrieval. Integration with External Data Sources: Connects with external data sources (e.g., weather reports, traffic data) to enrich the claim information. Enhances the accuracy of fraud detection by incorporating external context.

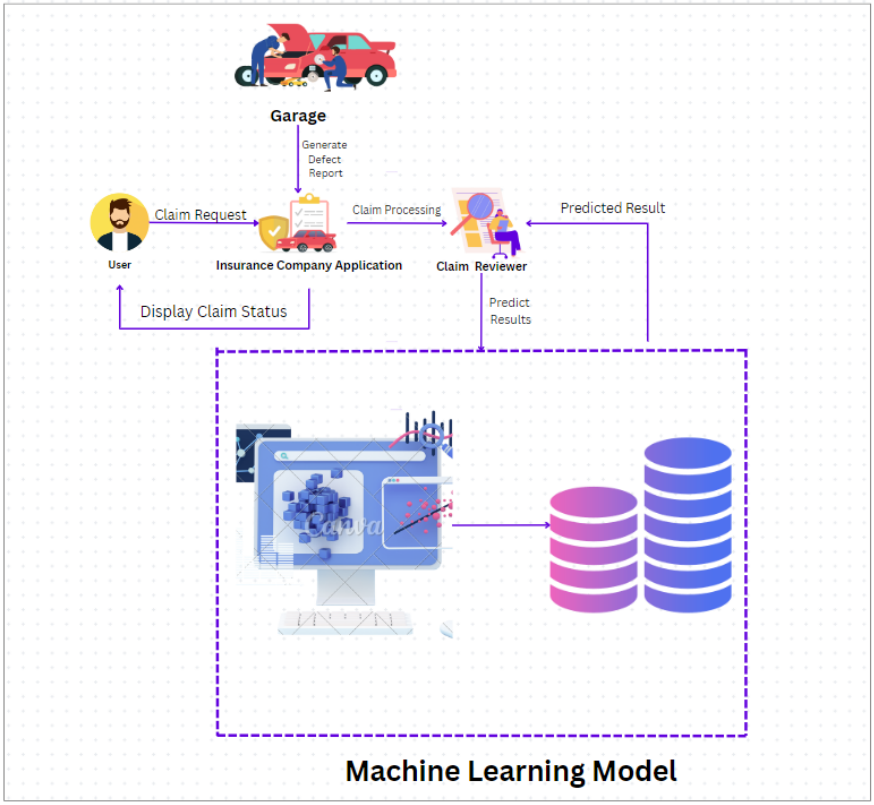


Figure 5.1: Insurance Fraud Detection System

2. **Claim Reviewer:**

   Purpose: The claim reviewer evaluates the submitted claims for initial assessment before passing them to the fraud detection model.

   Components and Functions:

   - **Claim Assessment Module:** Reviews basic claim information for validity. Flags claims with suspicious or inconsistent data for further investigation.
   - **Document Verification:** Verifies the authenticity of submitted documents (photos, police reports, etc.). Integrates Optical Character Recognition (OCR) for document analysis.
   - **Communication Module:** Facilitates communication with the insured party for additional information or clarification. Sends notifications to investigators if necessary.

3. **Investigator:**

   Purpose:

   Investigators handle in-depth analysis of claims flagged as suspicious by the initial review process.

   Components and Functions:

   - **Access to Claim Data:** Retrieves detailed claim data from the storage system. Accesses external data sources for additional context.
   - **Investigation Tools:** Provides tools for investigators to conduct thorough investigations. Includes data visualization, trend analysis, and anomaly detection tools.
   - **Communication Interface:** Enables communication with other stakeholders, such as claimants, law enforcement, or legal representatives. Logs communication history for reference.
   - **Fraud Indicators Identification:** Identifies potential fraud indicators based on patterns, historical data, and machine learning models. Generates reports summarizing investigation findings.

4. **Auto Insurance Fraud Claim Detection Model:**

   Purpose:

   This machine learning model is at the core of fraud detection, leveraging historical data and patterns to identify potentially fraudulent claims.

   Components and Functions:

   - **Feature Engineering:**

     Extracts relevant features from the claim data for model input. Utilizes domain knowledge to enhance feature selection.

   - **Machine Learning Model:**

     Trains and deploys a fraud detection model (e.g., supervised learning, anomaly detection).Constantly updates and retrains the model with new data for improved accuracy.

- **Integration with Claim Processing Flow:**
  Integrates seamlessly with the claim processing flow. Provides real-time predictions or flags for further investigation.
- **Explainability Module:**
  Incorporates an explainability module to provide insights into model predictions. Facilitates understanding and trust in the model by stakeholders.
- **Performance Monitoring:**
  Monitors the model's performance over time. Triggers retraining if the model's accuracy degrades.

5.2 DATA FLOW DIAGRAMS

A data-flow diagram is a way of representing a flow of data through a process of a system (usually an information system). The DFD also provides information about the outputs and inputs of each entity and the process itself. A data-flow diagram has no control flow.There are no decision rules and no loops. Specific operations based on the data can be represented by a flowchart.

Level 0:

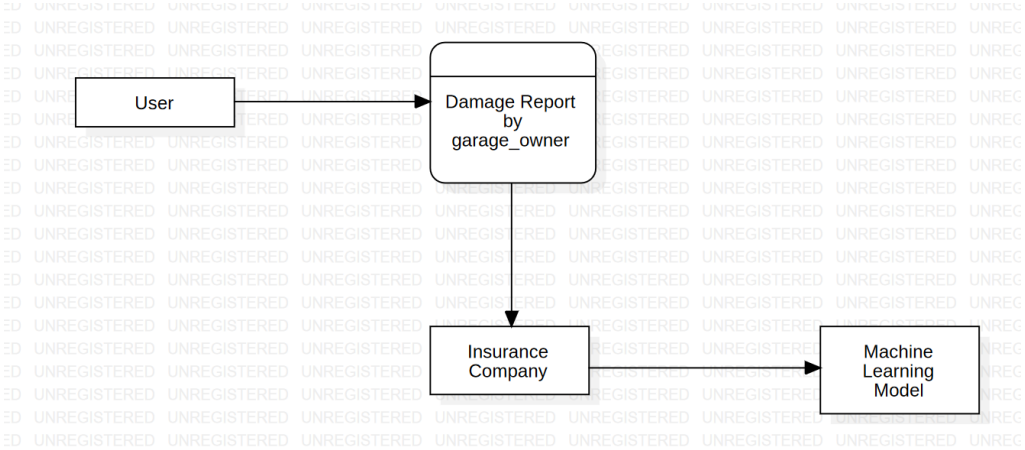


Figure 5.2: Data Flow Diagram(Level 0)

The Auto Insurance Claim Fraud Detection System utilizes a comprehensive Data Flow Diagram incorporating key stakeholders: users, garage owners, insurance

companies, and a sophisticated machine prediction model. Users play a pivotal role by initiating claims and furnishing essential information. Garage owners contribute valuable data by submitting information about vehicle damage and subsequent repairs. The insurance companies receive a stream of information from both users and garage owners, enabling a thorough evaluation of the legitimacy of each claim. At the heart of the system lies the machine prediction model, a cutting-edge technology employing advanced machine learning algorithms. This model meticulously analyzes historical data and discerns patterns indicative of potential fraudulent activities in auto insurance claims. Its role is not merely analytical but also strategic, offering profound insights to insurance companies. These insights empower them to make informed decisions regarding the authenticity of claims, thereby enhancing the overall efficiency and reliability of the auto insurance claim process. In essence, this collaborative framework seamlessly integrates human input with technological prowess to optimize fraud detection in the auto insurance domain[14]

Level 1:

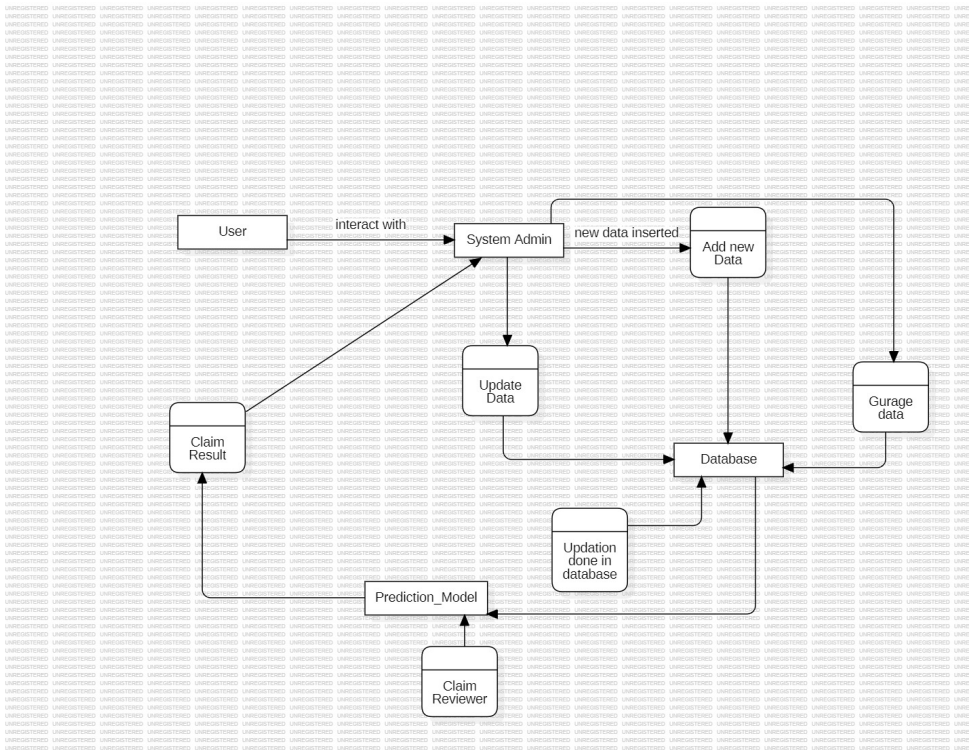


Figure 5.3: Data Flow Diagram(Level 1)

The data flow diagram (DFD) for the auto insurance fraud detection system revolves around key external entities and processes. External entities, such as Users, Garages, System Admins, and Claim Reviewers, interact with the system, contributing and validating data. The central processes include adding new data, updating system information, and the core operation of processing claims using a machine learning-based prediction model.

1. **Data Flow Processes:** Users and garages feed the system by adding new data, including insurance claims and damage assessments. System administrators ensure the database remains up-to-date through data updates. The pivotal "Process Claim" operation utilizes a machine learning prediction model to evaluate claim data for potential fraud. The results are then reviewed by claim reviewers, contributing to the system's ongoing refinement and accuracy.

2. **Data Flows and Storage:** Data flows from external entities to various processes, including "Add New Data" and "Update Data." Processed claim results, validated by claim reviewers, are stored in the database. This comprehensive approach involving external entities, core processes, and database management forms a robust framework for auto insurance fraud detection, emphasizing user interaction, continuous data updates, and reliable fraud prediction. This DFD design ensures a clear understanding of data flows, processes, and storage mechanisms, contributing to effective auto insurance fraud detection through machine learning [15].

3. **Data flow and System:** The Level 1 Data Flow Diagram (DFD) for vehicle insurance fraud detection using machine learning involves several key processes and entities. External entities include the Insurance Policyholder and the Insurance Company. The data flow begins with the Data Collection process, where relevant information is gathered from diverse sources such as policy details and claims history. Subsequently, the Data Preprocessing stage cleans and prepares the collected data for analysis, addressing issues like missing values. The Feature Extraction process identifies crucial features from the preprocessed data.

5.3 ENTITY RELATIONSHIP DIAGRAMS

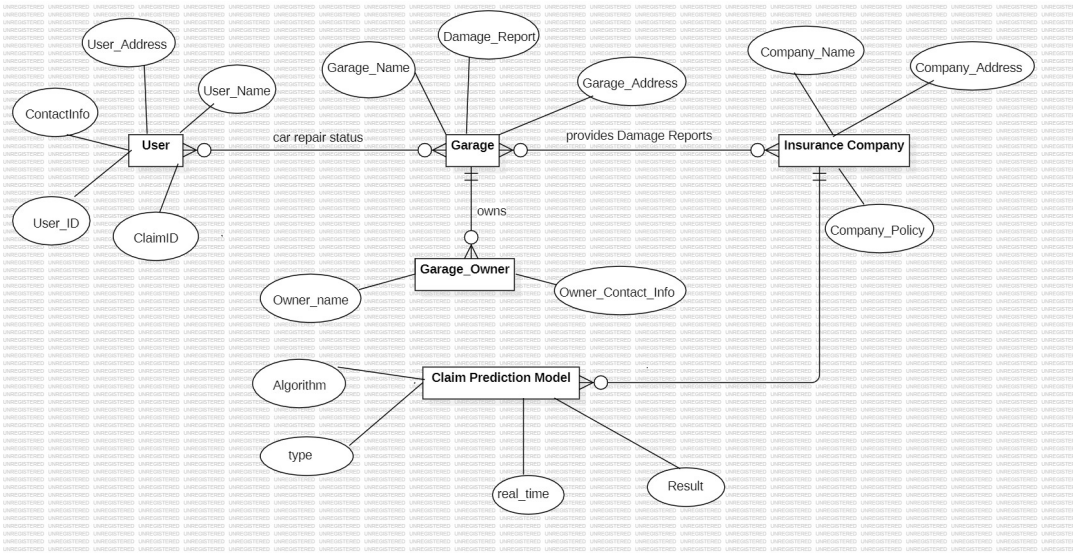


Figure 5.4: ER Diagram

An entity relationship diagram, also known as an entity relationship model, is a graphical representation that shows relationship among people, objects, concepts, or events within the information technology system.

1. **User:** This entity has attributes like UserAddress, ContactInfo , UserName, UserID and the id of the claim submitted i.e ClaimID The UserId is likely a unique identifier for each user. The ClaimID could be a regulatory requirement, and ContactInfo would include ways to get in touch with the Insurance company.

2. **Garage:** Garage plays an important role in overall ER diagram structure as the user visits after the accident to the garage . garage has the attributes GarageName, DamageReport, GarageAddress, which has many to many relationship with the User.

3. **GarageOwner:** It is another entity of the system which has the direct contact with the Insurance Company and provides the necessary requirement for the insurance company. And which has the many to one relationship with the garage.

4. **Insurance Company:** This entity has attributes like CompanyName, CompanyAddress, CompanyPolicy and the it has all the interaction with the Garage as well as the Claim Prediction Model entities and the type of relationship is given as the many to many with the Garage and the one to may with the Claim Prediction Model

5. **Claim Prediction Model:** This is the most important entity of the system and it has the attributes like Algorithm used , type of machine learning technologies used, real time fraud detection using the prediction statistics, and finaly the Result is the new attributes which gives the prediction about the claim , and these entity has the many to one relationship with the Insurance company entity.

5.4 UML DIAGRAMS

5.4.1 Class Diagram

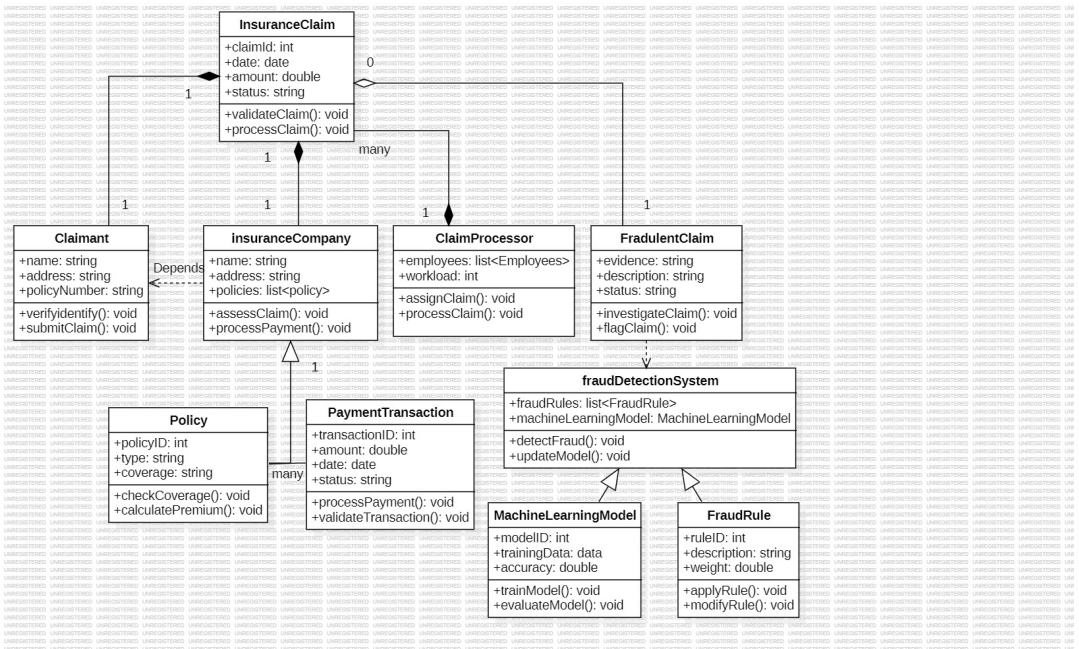


Figure 5.5: Class Diagram

A class diagram is a static diagram. It represents the application's static view. Class diagrams are used to create executable code for software applications as well as for

visualising, explaining, and documenting various elements of systems. The characteristics and functions of a class are described in a class diagram, along with the restrictions placed on the system. Fig. 5.5 represents basic class diagram.

1. **InsuranceClaim:**

   Attributes: claimID: string claimDate: date status: string amount: double description: string Relationships: Association with Claimant: Represents the relationship between an InsuranceClaim and the person making the claim (Claimant). Association with Policy: Indicates the association between an InsuranceClaim and the insurance policy to which it is related. Association with ClaimProcessor: Connects an InsuranceClaim to the entity responsible for processing the claim.

2. **Claimant:**

   Attributes: name: string address: string contactInfo: string policyID: string Relationships: Association with InsuranceClaim: Represents the claims submitted by a Claimant. Association with InsuranceCompany: Indicates the insurance company with which the Claimant has a policy.

3. **InsuranceCompany:**

   Attributes: companyID: string companyName: string location: string contactInfo: string Relationships: Association with Claimant: Represents the policies held by individuals with the InsuranceCompany. Association with ClaimProcessor: Connects the InsuranceCompany to the entity responsible for processing claims.

4. **ClaimProcessor:**

   Attributes: processorID: string processorName: string department: string contactInfo: string Relationships: Association with InsuranceClaim: Represents the claims processed by a ClaimProcessor. Association with InsuranceCompany: Indicates the insurance company employing the ClaimProcessor.

5. **FraudulentClaim:**

   Attributes: fraudID: string description: string investigationDetails: string Relationships: Association with InsuranceClaim: Represents a link between a

specific InsuranceClaim and a claim marked as fraudulent. Association with Investigator: Indicates the investigator responsible for handling fraudulent claims.

5.4.2 Usecase Diagram

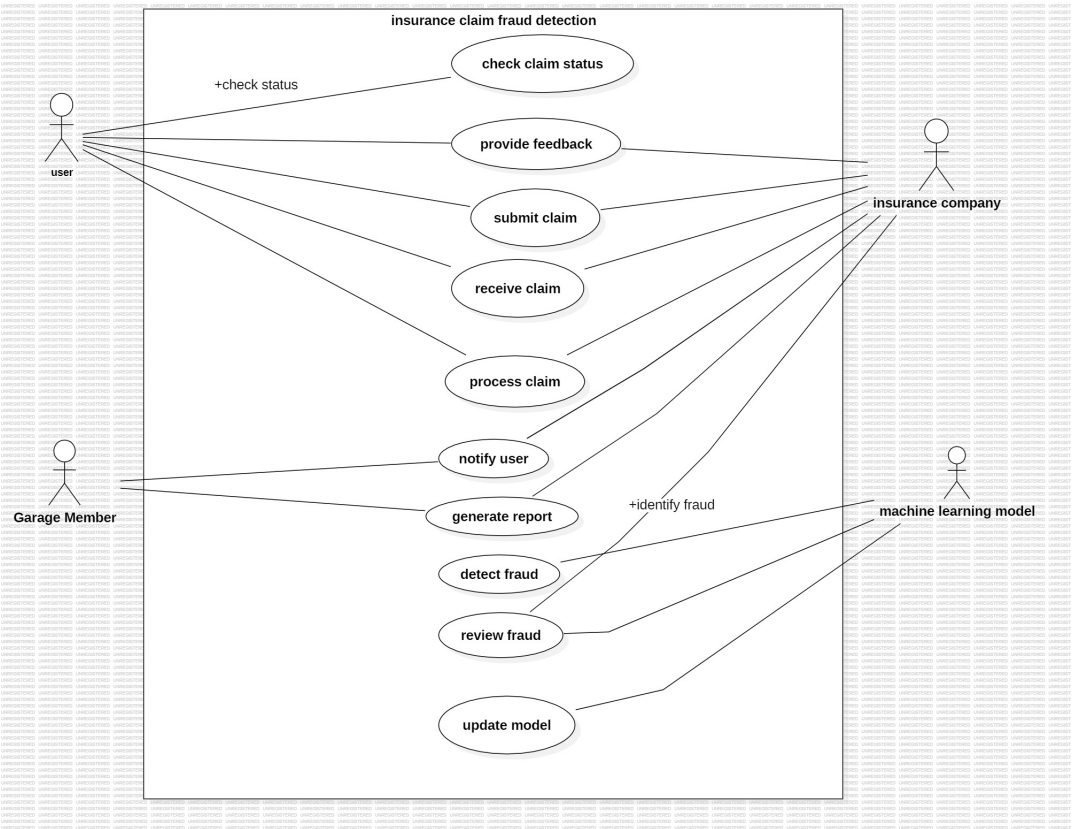


Figure 5.6: Usecase Diagram

1. **Check Claim Status:**

   Actor: User

   Description: Users can check the status of their auto insurance claims, providing transparency and updates on the progress of their requests.

2. **Provide Feedback:**

   Actor: User

   Description: Users have the ability to provide feedback on the claim process, enabling continuous improvement and addressing user concerns.

3. **Submit Claim:**

Actor: User

Description: Users initiate the claim process by submitting necessary information and documentation for assessment.

4. **Process Claim:**

   Actors: Insurance Company, Claim Prediction Model

   Description: Insurance companies process and assess claims, utilizing both human expertise and the predictive capabilities of the machine model to determine legitimacy.

5. **Notify User:**

   Actor: Insurance Company

   Description: The insurance company notifies users about the status and outcome of their auto insurance claims, ensuring clear communication throughout the process.

6. **Generate Report:**

   Actors: Insurance Company, Claim Prediction Model

   Description: The system generates comprehensive reports, combining human and machine analysis, to provide insights into the overall claim data and fraud detection results.

7. **Review Fraud:**

   Actors: Insurance Company, Claim Prediction Model

   Description: Insurance companies review potential fraud cases identified by the machine prediction model, making informed decisions based on both automated analysis and human expertise.

8. **Register:**

   Actor: User

   Description: Users can register in the system, providing necessary information for future claims and interactions with the auto insurance process.

   In the use case diagram for vehicle insurance fraud detection using machine learning, primary actors such as Users and Investigators are identified.

5.4.3 Activity Diagram

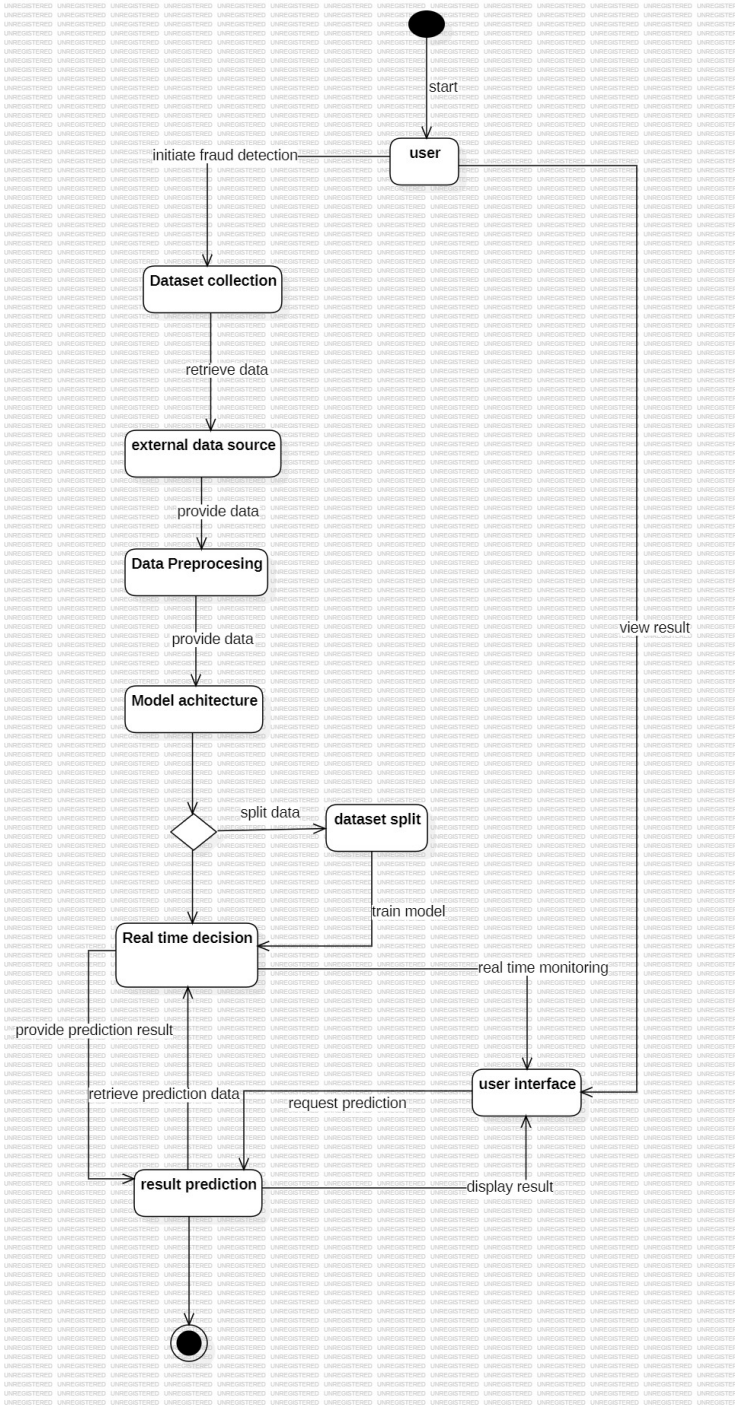


Figure 5.7: Activity Diagram

Fig 5.7 represent a very simplified, activity diagram is essentially a flowchart that shows how one activity leads to another. The action might be referred to as a system operation.One operation leads to the next in the control flow. This flow may be

parallel, contemporaneous, or branched. Activity diagrams use many features, such as fork, join, etc., to cope with all types of flow control. Similar to the other four diagrams, activity diagrams serve similar fundamental goals. It captures the system's dynamic behaviour. The message flow from one item to another is depicted using the other four diagrams, whereas the message flow from one activity to another is depicted using the activity diagram. An activity is a specific system function. Activity diagrams are used to build the executable system utilising forward and reverse engineering approaches, as well as to visualise the dynamic nature of a system [16].

Submit Claim Activity

1. **Claim Submission:** Actions: The Claimant initiates the claim submission process. Enters claim details (accident information, policy details, etc.). Uploads supporting documents. Claim Review Activity: Activity: Initial Claim Review Actions: Claim Processor reviews the submitted claim. Checks for completeness and validity.if necessary information is missing, requests additional details from the Claimant. Performs document verification using OCR. Decisions:

2. **Claim Investigation:** Actions: If the initial review flags the claim as suspicious, assign the claim to an Investigator. Investigator accesses detailed claim data. Utilizes investigation tools for in-depth analysis. Communicates with stakeholders for additional information.

3. **Fraud Detection** Actions: The system applies the auto insurance fraud claim detection model. Identifies potential fraud indicators based on historical data and machine learning models.

4. **Handling Fraudulent Claim:** Actions: Assign the fraudulent claim to an Investigator specializing in fraud cases. Investigator gathers additional evidence. Communicates with law enforcement if necessary. Generates a report on the fraudulent claim.

5. **Payment Processing:** Actions: If the claim is legitimate and not flagged as fraudulent, process the payment. Generate a payment transaction record. Update the claim status to "Closed."

6. **End Activity** Actions: End the claim processing workflow.

5.4.4 Sequence Diagram

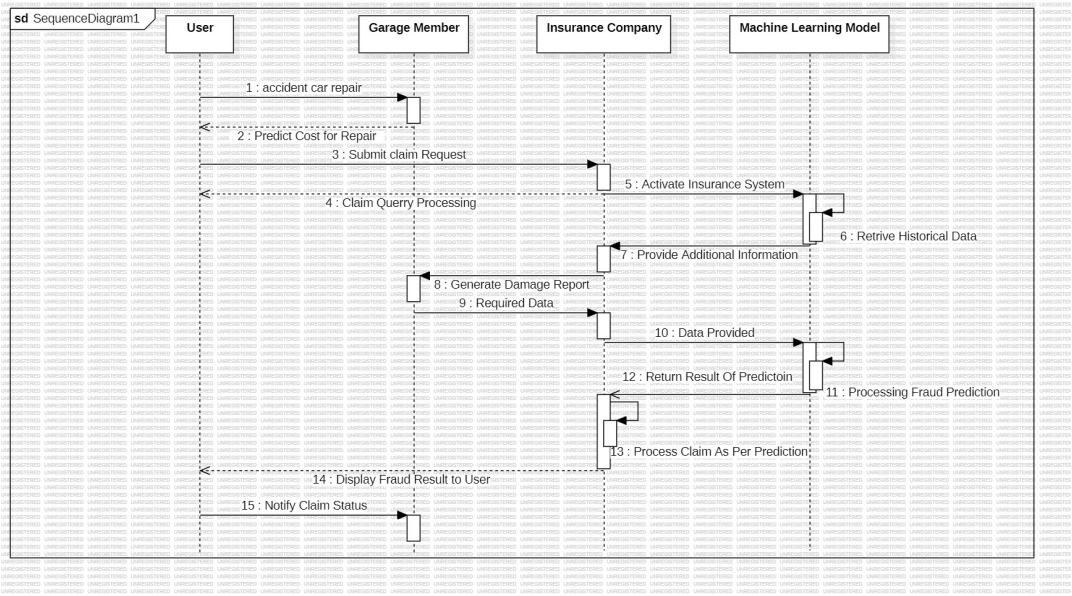


Figure 5.8: Sequence Diagram

Fig 5.8 represent a very simplified, Claim Submission SequenceThe Claimant initiates the claim submission.The system receives the claim submission request.The system validates the submitted data and documents.

1. **Claim Review Sequence:** The Claim Processor reviews the claim. If necessary information is missing, the system requests additional details from the Claimant. The system performs document verification using Optical Character Recognition (OCR).

2. **Investigation Sequence:** If the claim is flagged as suspicious, the system assigns the claim to an Investigator. The Investigator accesses detailed claim data. Utilizes investigation tools for in-depth analysis. Communicates with stakeholders for additional information.

3. **Fraud Detection Sequence:** The system applies the auto insurance fraud claim detection model. The model identifies potential fraud indicators based on historical data and machine learning models. The system returns the results of the fraud detection to the processing flow.

4. **Fraudulent Claim Handling Sequence:** If the fraud detection indicates potential fraud, the system initiates the fraudulent claim handling process. The system assigns the fraudulent claim to a specialized Investigator. The Investigator gathers additional evidence. Communicates with law enforcement if necessary. Generates a report on the fraudulent claim.

5. **Payment Processing Sequence:** If the claim is legitimate and not flagged as fraudulent, the system processes the payment. Generates a payment transaction record. Updates the claim status to "Closed."

6. **End Sequence:** The system concludes the claim processing workflow.

- **Key Elements in Sequence Diagrams:**

  - **Lifelines:** Represent entities or components involved in the sequence, such as Claimant, Claim Processor, Investigator, and the Fraud Detection Model.
  - **Messages:** Arrows represent messages exchanged between lifelines, indicating the flow of control or data.
  - **Activation Bars:** Represent the duration of an object's existence during a particular message exchange.
  - **Return Messages:** Indicate the flow of control back to the sender after the execution of a particular action.
  - **Focus of Control:** Use focus of control to highlight the current focus of execution within a lifeline. This sequence diagram provides a high-level view of the interactions and message flows in the insurance claim fraud detection process. It illustrates how different components collaborate and communicate to handle claims, conduct investigations, and detect potential fraud using machine learning techniques.

5.4.5 State Diagram

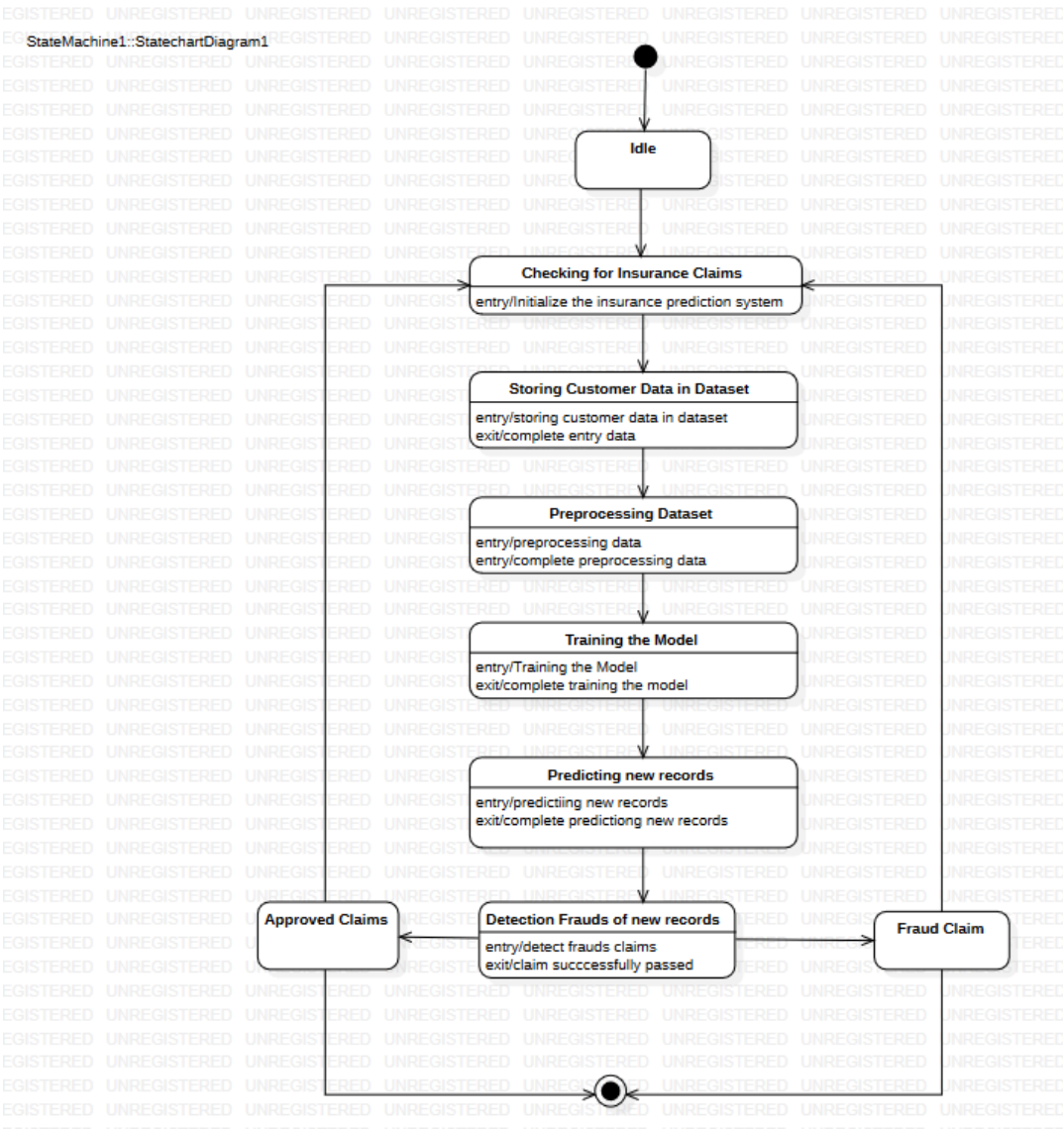


Figure 5.9: State Diagram

A state diagram, also known as a state machine diagram or statechart, is a type of behavioral diagram in the Unified Modeling Language (UML) that shows the states an object or an interaction can be in, as well as the transitions between those states. The arrows represent the transition from one state to another, and the descriptions on the arrows (entry, do, exit) represent actions taken when entering a state, while in the state, and exiting the state, respectively. Claim Review State: Description: Represents the state where the claim is reviewed by the initial claim processor.

Transitions:

1. Transition to ”Investigation” if the claim requires further scrutiny.

2. Transition to ”Fraud Detection” if the initial review indicates suspicious activity.

3. Transition to ”Payment Processing” if the claim is deemed legitimate.

Investigation State: Represents the state where an investigator is assigned to conduct an in-depth investigation. Transitions: Transition back to ”Claim Review” if the investigation clears the claim. Transition to ”Fraud Detection” if fraud indicators are identified.

Fraud Detection State: Represents the state where the system applies machine learning models to detect fraud. Transitions: Transition to ”Fraudulent Claim Handling” if fraud is detected. Transition back to ”Claim Review” if no fraud is detected.

Fraudulent Claim Handling State: Represents the state where a specialized investigator handles a detected fraudulent claim.

Payment Processing State: Represents the state where the system processes the payment for legitimate claims. Transitions: Transition to ”End” state after payment processing is complete.

End State: Represents the final state of the process. Transitions: No further transitions from this state. Additional Points: Initial State: The initial state is where the process begins, in this case, with the submission of an insurance claim. Final State: The ”End” state signifies the completion of the process. Transitions: Transitions represent the flow of the system from one state to another based on certain conditions or events.

Decision Points: Decision points, though not explicitly shown in a state diagram, can influence transitions between states. For example, the decision to transition from ”Claim Review” to ”Fraud Detection” depends on the outcome of the initial review.

Loops: Loops may exist within certain states, such as iterative investigations or reviews. This state diagram provides a high-level overview of the dynamic behavior of the insurance claim fraud detection system, emphasizing the states and transitions involved in processing claims and detecting potential fraud using machine learning techniques.

CHAPTER 6

PROJECT ESTIMATION, SCHEDULE AND TEAM STRUCTURE

Efficient software project estimating is a crucial and demanding task in the software development process. In the absence of a solid and trustworthy estimate, proper project planning and control are impossible. The software industry as a whole uses estimations improperly and performs a poor job of project estimating. As a result, we suffer considerably more than we ought to, and we should work to change the circumstances. Underestimating a project increases the likelihood of understaffing it, which can lead to staff fatigue, underspending on quality assurance efforts, which increases the chance of low-quality deliverables, and setting a schedule that is too short, which can damage the project's reputation by missing deadlines. Overestimating a project can be just as beneficial for those who plan to prevent this circumstance by liberally padding the estimate [**?**].

6.1 PROJECT SCHEDULE AND TEAM STRUCTURE

The four basic steps in software project estimation are:

1. Estimate the size of the development product. This generally ends up in either Lines of Code (LOC) or Function Points (FP), but there are other possible units of measure. A discussion of the pros and cons of each is discussed in some of the material referenced at the end of this report.

2. Estimate the effort in person-months or person-hours.

3. Estimate the schedule in calendar months.

4. Estimate the project cost in dollars (or local currency) Estimating size.

6.1.1 Estimating size

The first stage in creating an effective estimate is determining the size of the program that has to be produced with accuracy. If at all possible, you should begin your source(s) of knowledge on the project's scope with formal descriptions of the requirements, such as a system specification, software requirements specification, customer requirements specification, or request for proposal. Design documents can

be used to add more detail when [re-]estimating a project in its latter stages of development. You should still conduct a preliminary project estimate even in the absence of a written scope specification. Sometimes all you have to get started is a whiteboard outline or a verbal description. Regardless, you have to convey the degree of risk and unpredictability in an

6.1.2 Estimating effort

You may calculate the estimated effort after you have an idea of the product's size. Only with a clearly defined software development lifecycle and development process in place for the purpose of specifying, designing, developing, and testing the program can the conversion from software size to total project effort be completed. There is much more to a software development project than just writing the code; in fact, writing the code is frequently the least amount of the work involved. The majority of the project effort is spent creating and evaluating the deliverables, testing and reviewing the code, and creating and reviewing documentation. In order to create a project effort estimate, you must first identify, estimate, and then total all of the tasks that must be completed.

6.1.3 Estimating schedule

Determining the project schedule from the effort estimate is the third stage in software development project estimation. In most cases, this entails estimating the number of workers who will be assigned to the project, the tasks they will perform (the employment Breakdown Structure), as well as the start and end dates of their employment (the "staffing profile"). After obtaining this data, you must arrange it in a calendar timetable. Once more, you may utilize industry data models or historical data from previous projects at your company to estimate how many workers you'll need for a project of a specific size and how the labor will be divided into a timetable. When all else fails, use a schedule estimation rule. A Gantt chart is an effective tool for project management that shows a project's timetable graphically. The capacity of a Gantt chart to show a project's chronology makes it simple to comprehend the order of jobs and their duration, which is one of its main characteristics.

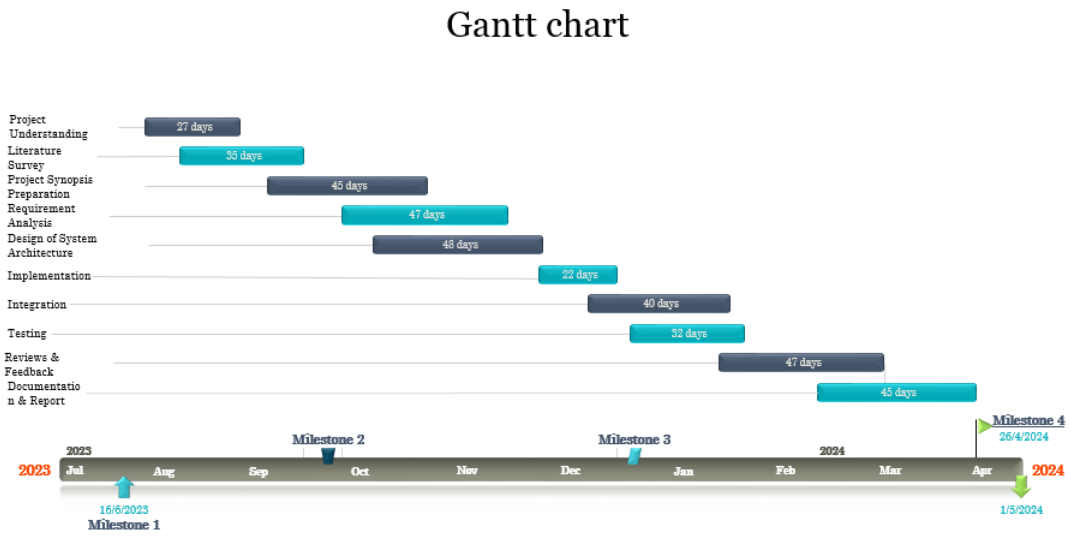


Figure 6.1: Gantt Chart

In a Gantt chart, jobs are arranged vertically down the side and the timeframe is shown horizontally. Every task is depicted with a bar that shows how long it will take to finish. The task's duration is indicated by the length of the bar, and its location on the timeline shows the task's start and end times.One advantage of representing projects with a Gantt chart is Gantt chart provides a comprehensive overview of a project's timeline, allowing project managers to plan, schedule, and track the progress of various tasks and activities. In addition to displaying the start and end dates of each task, Gantt charts often include additional information such as task dependencies, milestones, and resource assignments.

Task dependencies indicate the relationship between different tasks and help ensure that tasks are completed in the correct sequence. For example, if Task B cannot begin until Task A is complete, this dependency is represented in the Gantt chart to clearly illustrate the order of operations. A Gantt chart provides a comprehensive overview of a project's timeline, allowing project managers to plan, schedule, and track the progress of various tasks and activities. In addition to displaying the start and end dates of each task, Gantt charts often include additional information such as task dependencies, milestones, and resource assignments. Task dependencies indicate the relationship between different tasks and help ensure that tasks are completed in the correct sequence. For example, if Task B cannot begin until Task A is com-

plete, this dependency is represented in the Gantt chart to clearly illustrate the order of operations.

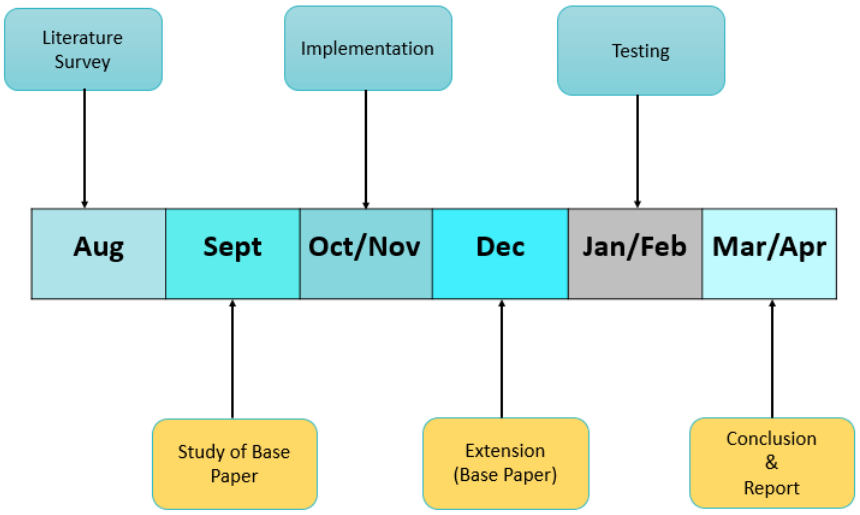


Figure 6.2: Schedule Estimation Chart

Milestones are important moments or occurrences in a project that signal a substantial accomplishment or the end of a stage. These benchmarks are easily recognized because they are frequently shown in the Gantt chart as diamonds or other unique symbols. Project teams can maintain focus on important goals and due dates with the aid of milestones. The resources or individuals of the team assigned to each task are indicated by their resource allocations. Project managers can make sure that work is distributed fairly and that team members are aware of their duties by adding resource assignments to the Gantt chart.

All things considered, Gantt charts are an invaluable tool for project managers since they offer a visual depiction of the project timetable, facilitating improved planning, scheduling, and progress monitoring. They aid in making sure that

6.2 PROJECT COST

When evaluating a project's overall cost, there are numerous aspects to take into account. These include of labor, the purchase or leasing of hardware and software, travel for meetings or testing, telecommunications (long-distance calls, video conferences, testing-dedicated lines, etc.), training sessions, office space, and so forth. The

precise method you use to estimate the total cost of the project will depend on how your company divides expenses. Certain expenses might not be attributed to specific projects; instead, they could be covered by increasing labor rates by an overhead value. A project manager for software development will frequently simply calculate the personnel cost and note any other project expenses that are not regarded as ”overhead” by the company. The project’s estimated effort (in hours) can be multiplied by to get the most basic labor cost.

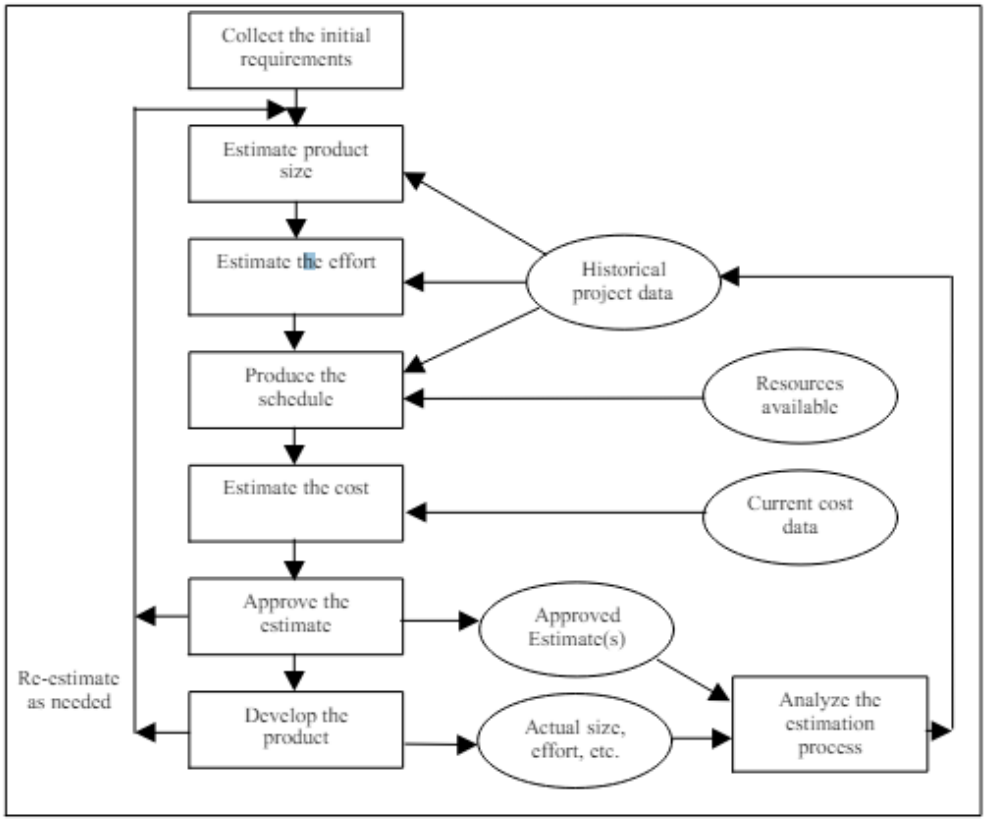


Figure 6.3: Project Cost Estimation Process

The project’s expected time is then determined by the amount of work needed. The resources’ capabilities and availability are taken into account when determining the project’s timetable. The predicted effort and time can be used to determine the project’s total cost. This entails taking other overhead expenses and the mean pay of the project’s staff members into account. The entire cost is divided into several areas, including staff, software and hardware, and other ancillary costs. An additional crucial phase in the cost estimation process is risk assessment.

CHAPTER 7

SOFTWARE TESTING AND VALIDATION

Determining the level of testing required for a machine learning-based car insurance claim fraud detection system is essential to guaranteeing its security, accuracy, and dependability. The quantity of testing necessary is determined by a number of variables, including risk tolerance, system complexity, and project needs. The degree of testing required is largely dependent on how vital the system is. To reduce the chance of false positives and false negatives, thorough testing is required if the fraud detection system is mission-critical. Comprehensive testing is essential since these mistakes could have serious financial repercussions for the insurance provider. Furthermore, more thorough testing to guarantee data security and privacy could be required in order to comply with regulatory standards like GDPR or HIPAA. Adherence to these guidelines is crucial in safeguarding confidential data.

7.1 TYPE OF TESTING

Implementing various types of testing is essential for developing a robust car insurance claim fraud detection system using machine learning.

7.1.1 Unit Testing:

This involves testing individual components or units of the system in isolation. For instance, unit testing can be applied to assess the accuracy and functionality of machine learning models, algorithms, or specific functions responsible for data preprocessing, feature extraction, and fraud detection.

7.1.2 Integration Testing:

Integration testing verifies the interaction between different modules or components of the system. In the context of the fraud detection system, it ensures that various machine learning models, data processing pipelines, and other system components work together seamlessly.

7.1.3 Functional Testing:

This evaluates the system's functionality against the specified requirements. In the case of the fraud detection system, functional testing ensures that the system accu-

rately detects fraudulent insurance claims according to predefined criteria, meeting business requirements effectively.

7.1.4 Performance Testing:

This assesses the system's responsiveness, scalability, and stability under various conditions. For the fraud detection system, performance testing measures its ability to handle large volumes of insurance claims data efficiently and provide timely fraud detection results.

7.1.5 Accuracy Testing:

It evaluates the accuracy of the machine learning models used in the fraud detection system by comparing the system's predictions against known outcomes. This ensures the system's effectiveness in accurately detecting fraudulent insurance claims.

7.1.6 Robustness Testing:

Robustness testing assesses the system's ability to handle unexpected inputs or invalid data gracefully. For the fraud detection system, robustness testing evaluates how well the system performs when faced with outliers, noise, or other irregularities in the insurance claims data.

7.1.7 Security Testing:

Security testing verifies the system's ability to protect sensitive data and prevent unauthorized access. For the fraud detection system, security testing ensures that the system complies with data protection regulations and safeguards the confidentiality and integrity of the insurance claims data.

7.1.8 Usability Testing:

Usability testing evaluates the system's user interface and user experience. For the fraud detection system, usability testing assesses the system's ease of use and ensures that users can interact with the system effectively to review and validate the fraud detection results.

7.2 TEST CASE

Table 7.1: Test Cases for Vehicle Insurance Fraud Detection System

ID	Test Case Description	Inputs	Expected Output
TC-01	Data Collection Verification	Dataset of vehicle insurance claims	Successful retrieval and preprocessing of data
TC-02	Feature Engineering Validation	Raw data attributes and preprocessing techniques	Extracted and engineered features ready for model input
TC-03	Model Training Check	Preprocessed data and model training parameters	Trained model with optimized parameters
TC-04	Model Evaluation	Test dataset and trained model	Evaluation metrics meeting predefined thresholds
TC-05	Performance Analysis	Classification metrics and model performance indicators	High accuracy, precision, recall, and F1 Score
TC-06	Class Imbalance Handling	Imbalanced dataset and SMOTE technique application	Balanced dataset with improved model performance
TC-07	Hyperparameter Tuning	Model architecture and hyperparameter optimization details	Improved model performance after tuning
TC-08	Model Interpretability	Feature importance analysis and model interpretation	Clear understanding of factors influencing predictions
TC-09	Real-world Scenario Testing	Simulated insurance claims and model inference	Accurate detection of fraudulent claims in real-time
TC-10	Integration Testing	Integration of model with insurance company infrastructure	Seamless integration and functionality of the system

7.3 RISK MANAGEMENT

The process of creating a machine learning-based system for detecting fraudulent auto insurance claims requires careful consideration of risk assessment and analysis. Project teams can lessen the possibility of expensive delays or failures by proactively addressing and mitigating potential risks and evaluating their potential impact. In order to identify any potential risks to the project's success, the project team thoroughly reviews every facet of the undertaking during the risk identification phase. The development of an automobile insurance claim fraud detection system is often

fraught with dangers, such as problems with data availability and quality, model performance, regulatory compliance, scalability, and security vulnerabilities. Following their identification, possible risks are examined to determine how likely they are to occur and what effect they might have on the project.

Lastly, in order to lessen the impact of hazards that have been recognized, risk mitigation methods are created. This can entail putting in place extra security measures, enhancing data quality procedures, or creating backup plans. Project teams can prevent problems before they arise and guarantee the effective implementation of the vehicle insurance claim fraud detection system by methodically identifying and evaluating possible hazards. The methodical process of risk identification is essential to the creation of a machine learning-based system for detecting fraudulent auto insurance claims. It entails determining possible dangers that might impede the system's effectiveness. Issues with data availability and quality, model performance, regulatory compliance, scalability, and security vulnerabilities are examples of common hazards. The accuracy of the fraud detection system can be greatly impacted by incomplete or poor-quality data, and regulatory non-compliance may result in legal consequences.

Once potential risks are identified, they undergo analysis to assess their likelihood and potential impact. Risk probability is determined by evaluating the likelihood of each risk occurring, while risk impact assesses the potential consequences on the project's objectives, schedule, budget, and quality. Risks are then prioritized based on their probability and impact, allowing project teams to focus their efforts on addressing the most critical risks first. High probability/high impact risks are given the highest priority, followed by medium and low-risk categories.

CHAPTER 8

RESULT AND ANALYSIS

The project assessed categorization performance using a confusion matrix, as indicated in Table I. Important metrics like accuracy, sensitivity (or true positive rate), and specificity were based on this matrix. While sensitivity means the ratio of rightly identified positive claims to the total number of real legitimate claim, accuracy evaluates the overall soundness of model's predictions. The percentage of correctly diagnosed negative cases relative to all negative cases is known as specificity. These metrics provide critical insights into the model's capacity to distinguish between different classes, which is necessary for assessing its practical usefulness. Using the confusion matrix and accompanying metrics, the study acquires a thorough insight into the model's performance, allowing for informed judgments about its deployment and optimization in real-world applications. In addition to accuracy, sensitivity, and specificity, the confusion matrix also provided other important performance metrics such as precision, F1 score, and the area under the receiver operating characteristic curve (AUC-ROC).

Precision measures the proportion of true positive predictions out of all positive predictions made by the model. It indicates the model's ability to avoid false positives. The F1 score, which is the harmonic mean of precision and sensitivity, provides a balanced measure of the model's performance, particularly in situations where there is an imbalance between the classes. The AUC-ROC, on the other hand, measures the model's ability to distinguish between the positive and negative classes across all possible thresholds. A higher AUC-ROC value indicates better overall performance of the model. By analyzing these metrics in conjunction with the confusion matrix, the study was able to comprehensively evaluate the model's performance, providing valuable insights into its strengths and weaknesses. These insights are essential for making informed decisions regarding the model's deployment and optimization in real-world applications. Confusion matrix is essential tool for assessing a measuring the classification of model's based on performance in the confusion matrix. Offering a wide variety of entries in tabular

overview of the model's predictions in comparison to the actual labels. Usually, the matrix has four entries:

1. True Positive (TP): Cases where the model rightly identified as right.

2. True Negative (TN): Examples that the model rightly categorized as wrong.

3. False Positive (FP): Cases when the model misclassified something as positive.

4. False Negative (FN): Examples of data that the model misclassified as negative.

The classifier's parameter evaluation is displayed in Table II. The analysis the performance metrics derived from the provided table, revealing distinct patterns among four machine learning algorithms: Random Forest Outperforms other algorithms Decision Tree Building numerous trees, AdaBoost basically an Ensemble technique, and Support Vector Machine (SVM) a simple approach used to solve complex classification problems. Notably, Random Forest emerges as a top performer. Following Result Comparison table shows with 4663 instances correctly classified and 1136 incorrectly classified, alongside a precision of 0.8923, recall of 0.7647, and an F1 Score of 0.8236. Decision Tree closely follows, achieving 4635 correct classifications, 1164 misclassifications, and precision, recall, and F1 Score of 0.8876, 0.7607, and 0.8193, respectively. Conversely, AdaBoost demonstrates lower accuracy, with 3405 correct classifications, 2393 misclassifications, and a precision of 0.4591, recall of 0.6348, and F1 Score of 0.5328. SVM, too, exhibits relatively poorer performance, correctly classifying 3231 instances, misclassifying 2568, and attaining a precision of 0.3504, recall of 0.6206, and F1 Score of 0.4479. These numerical insights underscore Random Forest and Decision Tree's superiority over AdaBoost and SVM in classification tasks, emphasizing their higher accuracy, precision, and F1 Score. The information of the below table gives the useful insight of the performance of the machine learning models; results may be interpreted as needed.

Table 8.1: Model Performance Metrics

Metric	**Random Forest**	**Decision Tree**	**AdaBoost**	**SVM**
Correctly classified	4663	4635	3405	3231
Incorrectly classified	1136	1164	2393	2568
Precision	0.8923	0.8876	0.4591	0.350
Recall	0.7647	0.7607	0.6348	0.620
F1 Score	0.8236	0.8193	0.5328	0.447
Accuracy	80%	78%	60%	58%

Given Metrics gives various term which are When working on assignments when there is an imbalance between the classes or when it's necessary to strike a balance between false positives and false negatives, these metrics are especially important. There are various terminologies helpful for analysis of the model can be

derived using the confusion matrix which are as follows:

1) Precision Precision is the ratio of true positives to the total predicted positives. It measures the accuracy of positive predictions. The formula for precision is True Positives divided by the sum of True Positives and False Positives. It provides insight into the precision of positive forecasts.

2) Accuracy Accuracy represents the proportion of correctly anticipated observations to the total observations, indicating the model's overall correctness. When the dataset's class distribution is balanced, accuracy is a valuable statistic. It may, however, be deceptive when there is a disparity in class.

3) F1 Score When the dataset's class distribution is balanced, accuracy is a valuable statistic. It may, however, be deceptive when there is a disparity in class.

Where P is precision, R is recall. When there is an imbalance in the classes, the F1 score is especially helpful. Both false positives and false negatives are taken into account. A comparison was conducted between four machine learning algorithms: RF, DT, AdaBoost, and SVM Classifier. These algorithms are frequently used in classification jobs with the goal of precisely identifying fraudulent claims. With an accuracy of 80method performed better than the other algorithms. This implies the Random Forest makes strong forecasts by efficiently utilizing the joint judgments of several decision trees. A basic algorithm called Decision Tree obtained an accuracy of 78Forest. Decision Tree is remarkably effective in this assignment, even if it is simple. AdaBoost, a boosting method, performed relatively worse than the prior algorithms, with an accuracy of 60circumstances by iteratively adjusting the weights of misclassified examples, yet it was unable to match the performance of Random Forest and Decision Tree in this scenario. SVM Classifier, a method based on defining hyperplanes to separate classes, demonstrated the lowest accuracy among the four algorithms, at 58

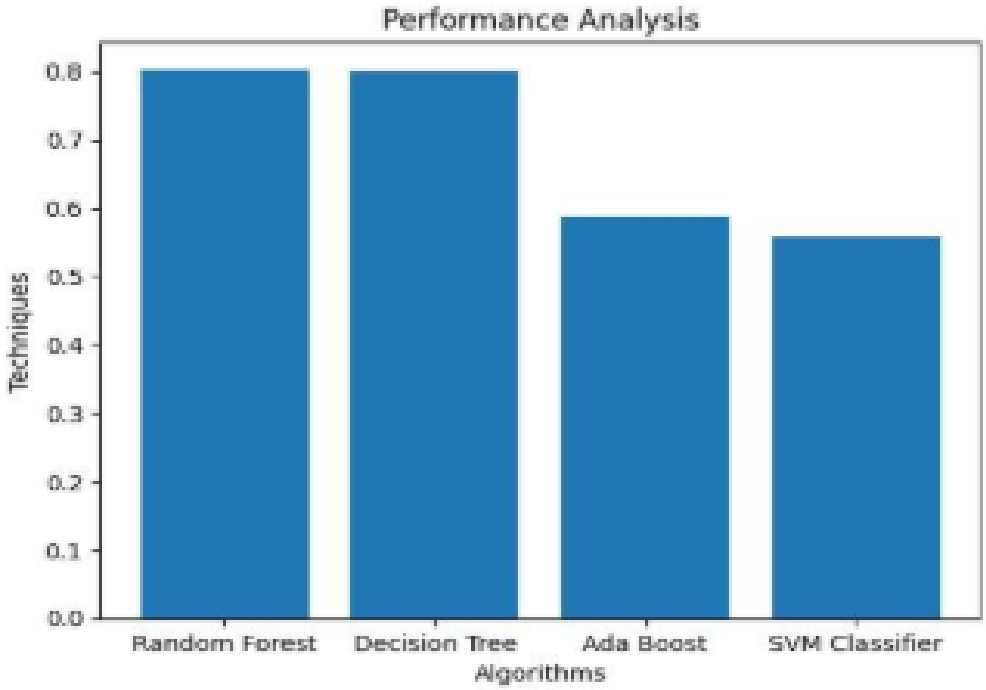


Figure 8.1: Performance Analysis

Both the insurance company and the customers are quite concerned about fraudulent claims on auto insurance. Application of anomaly detection in machine learning can be achieved using above methodology. Numerous research projects have used several machine learning algorithms for detecting fake information. It proposes to use RF, DT C4.5 as a supervised classifier to differentiate between legitimate and fraudulent claims. In order for the training data to generate a sufficiently accurate model, SMOTE must be suggested. Performance of model is tested using testing data, a highly skewed dataset that represents real-world data. The result shows that RF, DT C4.5, and all of them attain good accuracy. Nonetheless, RF performs the best with 89.23

8.1 IMPLEMENTATION

The homepage of the auto insurance claim fraud detection system using machine learning serves as the main interface for users, providing them with essential tools to combat fraudulent claims efficiently. Upon accessing the system, users are prompted to log in or register for an account if they are new users. Once logged in, users

are greeted with a comprehensive navigation menu, offering easy access to various sections of the system. These sections typically include a dashboard, claim analysis, model performance, and settings.

The dashboard section of the homepage offers users a quick overview of recent activities and system insights. It provides summary statistics of processed claims, real-time updates on new claims, and their current status, as well as alerts for potential fraudulent claims. This feature allows users to stay informed and take immediate action when necessary.

In the claim analysis section, users can delve into the details of individual insurance claims. They can access comprehensive information about claimants, claim amounts, and assessment results. Additionally, this section provides users with visualizations and insights derived from the claim data, enabling them to make informed decisions regarding the authenticity of each claim.

The model performance section allows users to review the performance metrics of the machine learning models used for fraud detection. Here, users can access metrics such as accuracy, precision, recall, and F1 score, as well as confusion matrices and ROC curves. This information helps users evaluate the effectiveness of the models and make any necessary adjustments for improved performance.

Finally, the settings section of the homepage enables users to customize their preferences. They can manage notification settings, update their profile information, and configure system settings according to their specific requirements. Overall, the homepage of the auto insurance claim fraud detection system provides users with the necessary tools and information to effectively combat fraudulent claims and ensure the integrity of insurance processes. New users are guided through a straightforward registration process, where they can create an account by providing necessary details and credentials. This ensures that the system maintains a database of authorized users, enhancing security and accountability.

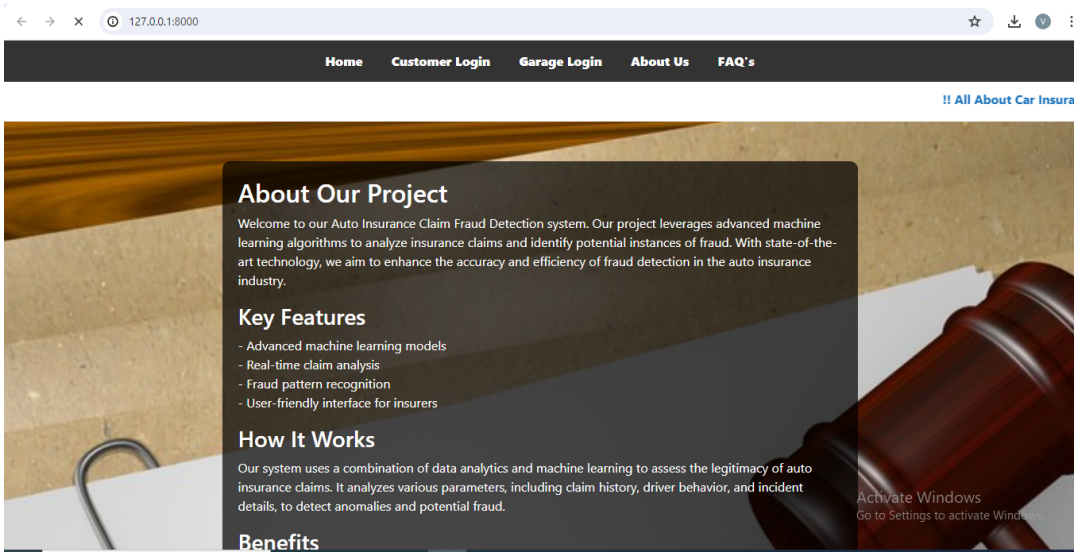


Figure 8.2: Home Page

The login page of the auto insurance claim fraud detection system using machine learning serves as the initial gateway for users to access the system's functionalities. It provides a secure and user-friendly interface where registered users can authenticate their identities and gain access to the system. New users are guided through a simple registration process to create an account, ensuring that only authorized individuals can interact with the system.

Once on the login page, users are prompted to enter their credentials, typically consisting of a username and password. These credentials are verified against the system's database to ensure the user's identity and grant access to the system's features. For added security, the login page may also incorporate additional authentication measures, such as multi-factor authentication or CAPTCHA verification, to prevent unauthorized access.

Upon successful authentication, users are redirected to the system's homepage, where they can access a range of features and tools to analyze insurance claims, evaluate model performance, and customize their preferences. The login page thus serves as the initial point of entry for users, ensuring secure access to the auto insurance claim fraud detection system and providing a seamless user experience.

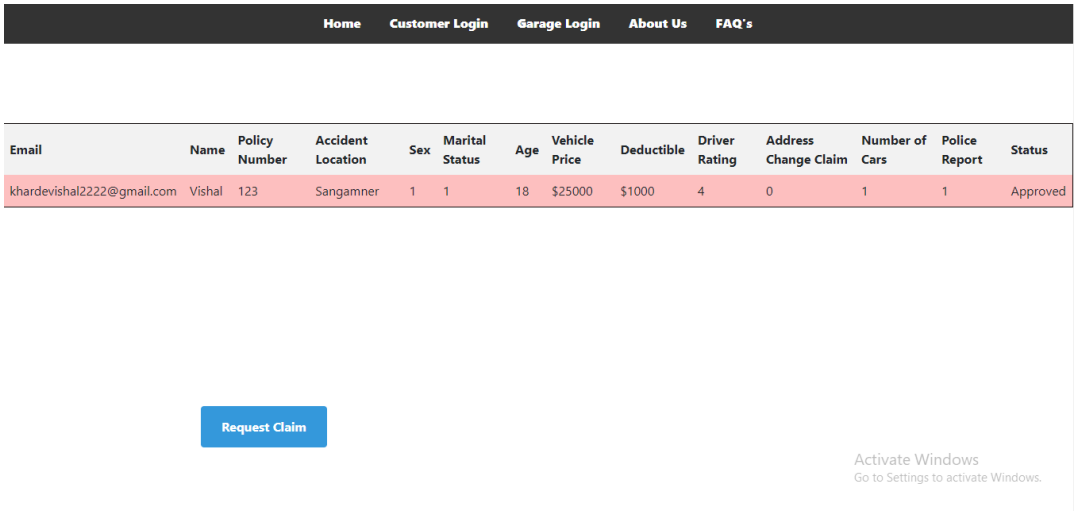


Figure 8.3: Customer Login Page

The login page features a user-friendly interface where registered users can enter their credentials, typically comprising a username and password. These credentials are securely verified against the system's database to authenticate the user's identity. To further enhance security, the login page may incorporate additional authentication measures such as multi-factor authentication or CAPTCHA verification.

Upon successful authentication, users are granted access to the system's homepage, where they can explore a range of features and tools. These include dashboard analytics, claim analysis, model performance evaluation, and settings customization. By providing a secure and user-friendly login experience, the system ensures that users can efficiently leverage machine learning technologies to detect and prevent auto insurance claim fraud. The login page of the auto insurance claim fraud detection project serves as the entry point for authorized users to access the system's functionalities. It plays a critical role in ensuring the security and privacy of the system by authenticating users before granting access.

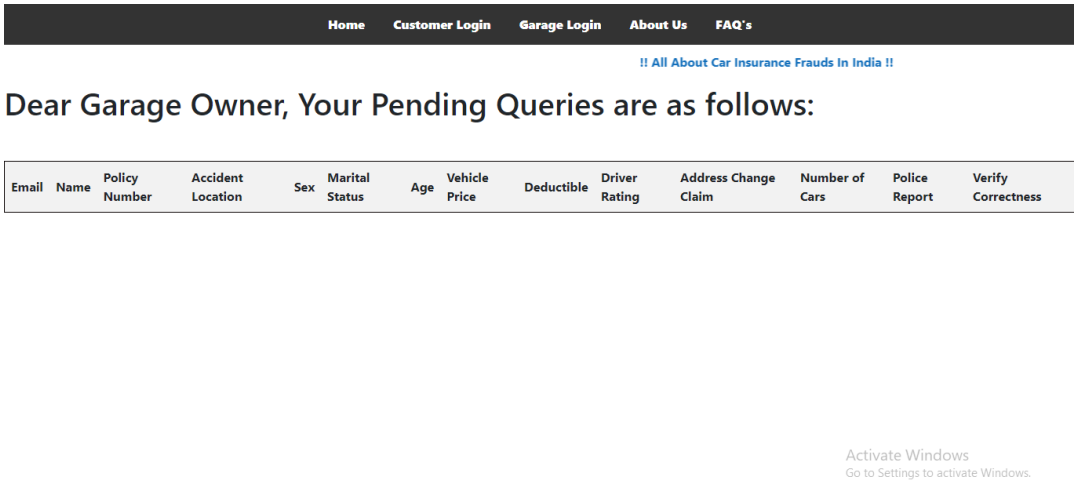


Figure 8.4: Garage Page

Upon successful authentication, users are redirected to the system's home-page, where they can access a range of features and tools to analyze insurance claims, evaluate model performance, and customize their preferences. The login page thus serves as the initial point of entry for users, ensuring secure access to the auto insurance claim fraud detection system and providing a seamless user experience.

CHAPTER 9

ADVANTAGES, LIMITATIONS AND APPLICATION

9.1 ADVANTAGES

Using machine learning for insurance fraud claim detection can offer several advantages, enhancing the efficiency and accuracy of the process. Here are some key advantages:

1. **Early Detection:** Machine learning algorithms can analyze large volumes of data quickly and identify patterns that may indicate fraudulent activity. This enables early detection of potential fraud, allowing insurers to take proactive measures.

2. **Improved Accuracy:** Machine learning models can be trained on historical data to learn patterns and anomalies associated with fraudulent claims. This can lead to more accurate detection compared to traditional rule-based systems, which may be less adaptive to evolving fraud tactics.

3. **Reduced False Positives: :** By incorporating advanced analytics, machine learning models can be fine-tuned to reduce false positives. This means legitimate claims are less likely to be flagged as fraudulent, improving customer satisfaction and reducing unnecessary investigations.

4. **Automation and Efficiency:**

   Machine learning can automate the claims processing workflow, saving time and resources for insurance companies. Automation can enable decisions to be made faster and allow customers to focus on more complex tasks.

5. **Make improvements to change strategies:** Scammers continue to evolve their strategies, making it difficult to detect fraud patterns using static rules. But machine learning models can adapt to new fraud schemes by learning from new data, making them more resilient to emerging threats.

6. **Data integration:** Machine learning systems can integrate and analyze different data sources, including structured and unstructured data. This comprehensive analysis provides a broader view of the claims and helps identify subtle patterns that indicate fraud.

7. **Cost Savings:** Insurance companies can reduce fraud costs by streamlining and streamlining fraud detection processes. Detect and handle fraudulent applications. This allows greater utilization of resources and greater control over payments.

8. **Effective fraud detection:** Machine learning algorithms can improve over time as they continue to learn from new data. This means that fraud patterns are becoming more complex and accurate as we encounter and adapt to new types of fraud.

9.2 LIMITATIONS

Although machine learning has significant advantages in insurance fraud detection, its application also has some limitations and challenges. Some important limitations are:

1. **Dataset Imbalance:** In insurance fraud investigation, the number of fraud cases is often lower than the data claimed. Legally, this makes the data unreliable. This can lead to a biased sample for most of the group and make it difficult to detect cases of fraud.

2. **Data Quality:** The effectiveness of machine learning models depends on the quality and completeness of the data they use. Education. Incorrect or missing data can lead to inaccurate models and poor performance.

3. **The nature of fraud:** Fraudsters are constantly evolving their strategies, making the job of traditional machine learning models difficult. It will be difficult for patterns learned from historical data to detect new and changing fraud patterns.

4. **False Positives and False Positives:** Machine Learning Models can produce false positives to deceive, prove true as false positives, or produce false positives, false positives, or false positives. Striking the right balance between sensitivity and specificity is difficult.

5. **Interpretability:** Many machine learning models are uninterpretable, especially complex models such as deep neural networks. It can be difficult to understand how the model reaches certain decisions, which can prevent those decisions from being explained to stakeholders, management, or customers.

6. **Hostile Attacks:** Scammers may attempt to use methods to deliberately provide misleading information to evade detection. Machine learning models can be vulnerable to attacks where fraudsters use flaws in the model to fool them.

7. **Integration Discussion:** Implementing machine learning models into existing systems and workflows can be challenging. Integration issues may arise due to differences in input data, data storage, or other assumptions.

8. **Ethical considerations:** Using machine learning in fraud detection raises ethical issues such as privacy concerns and may cause discrimination issues. It is important to ensure that algorithms are fair, fair and not biased against certain groups of people.

9.3 APPLICATIONS

The application of machine learning in insurance fraud claim detection is diverse and can be applied across various stages of the claims process. Here are some key applications:

1. **Claims Triage:** Machine learning algorithms can be used to automatically triage incoming claims based on their likelihood of being fraudulent. This helps insurance companies prioritize high-risk claims for further investigation, ensuring that limited resources are directed toward cases with the highest potential for fraud.

2. **Anomaly Detection:** Machine learning models can analyze historical claims data to identify patterns and establish a baseline for what constitutes normal behavior. Any deviation from this baseline can be flagged as an anomaly, potentially indicating fraudulent activity.

3. **Pattern Recogition:** By analyzing large datasets, machine learning can identify patterns and trends associated with known fraud schemes. This includes identifying common characteristics, behaviors, or relationships among claims that may indicate fraudulent activity.

4. **Text and Sentiment Analysis:** Natural language processing (NLP) techniques can be employed to analyze text data in claims forms, correspondence, or other documents. Sentiment analysis can help detect inconsistencies or suspicious language that may be indicative of fraud

5. **Social Network Analysis:** Machine learning can be used to analyze the relationships between different entities, such as claimants, policyholders, and service providers. Uncovering complex networks of relationships can reveal potential collusion or fraud rings

6. **Predictive Modeling:**

   Predictive modeling techniques, such as logistic regression or decision trees, can be used to assess the likelihood of a claim being fraudulent based on a combination of features. These models can be trained on historical data and updated as new information becomes available.

SUMMARY AND CONCLUSION

In summary, the development of a machine learning-based auto insurance claim fraud detection system represents a significant step forward in addressing the challenges posed by insurance fraud. By leveraging advanced analytics and artificial intelligence, insurers can enhance their ability to detect and prevent fraudulent activities, thereby safeguarding their financial interests and maintaining the trust of their customers. This project provides a comprehensive framework for building such a system, with the potential to deliver tangible benefits to both insurers and policyholders alike.

Both the insurance company and the customers are quite concerned about fraudulent claims on auto insurance. Application of anomaly detection in machine learning can be achieved using above methodology. Numerous research projects have used several machine learning algorithms for detecting fake information. It proposes to use RF, DT C4.5 as a supervised classifier to differentiate between legitimate and fraudulent claims. In order for the training data to generate a sufficiently accurate model, SMOTE must be suggested. Performance of model is tested using testing data, a highly skewed dataset that represents real-world data. The result shows that RF, DT C4.5, and all of them attain good accuracy. Nonetheless, RF performs the best with 89.23accuracy. Such habit may be learned once and for all, or it may continue to change over time. Regression, classification, and clustering are the additional classifications for the supervised learning. Predicting which class an observation belongs to is the definition of classification; clustering divides the observation into meaningful groups. Regression allows us to forecast value based on observation . Assigning a document to a predetermined category is the fundamental notion of classification. Data and a program are the inputs for the conventional method . The computer receives the inputs and produces the output in the end. When it comes to machine learning, input and output go into a computer, which then produces a program as an output. Below is a comparison of the machine learning approach and the classical learning strategy. the development of a machine learning-based auto insurance claim fraud detection system holds immense promise in the realm of financial security, risk management, and customer trust within the insurance industry. This project has aimed to address the critical issue of fraudulent claims, which not

only cause substantial financial losses but also erode the trust between insurers and their customers. Through an in-depth exploration of the problem, literature, methodology, and potential outcomes, this project sets out a comprehensive roadmap for building an effective fraud detection system. The research began by recognizing the pressing need for advanced fraud detection methods due to the significant financial impact of insurance fraud, estimated to be in the billions of dollars annually. The existing systems were found to be inefficient and costly, prompting the exploration of machine learning as a solution. By leveraging real-world data from an automobile insurance company in Indonesia, the study aimed to evaluate the effectiveness of various machine learning algorithms for fraud prediction. The results indicated that Random Forest performed best among the algorithms tested, showcasing its potential for enhancing fraud detection accuracy. Through a detailed exploration of related literature, this project gained insights into existing methodologies and approaches in fraud detection across different domains, such as health insurance and general insurance. These studies provided valuable guidance on algorithm selection, feature engineering, and model evaluation. They also highlighted the significance of adapting to emerging fraud tactics and the importance of continuous learning to maintain model effectiveness. The proposed methodology involved several key modules, including dataset collection and preprocessing, model architecture, real-time fraud detection, and user interface development. Each module plays a crucial role in building a robust fraud detection system. Data preprocessing ensures the quality and reliability of the dataset, while model architecture and real-time detection contribute to the accuracy and efficiency of fraud identification. The user interface module enhances usability and accessibility, allowing stakeholders to interact with the system effectively. Moreover, the project outlined the software and hardware requirements necessary for system implementation, emphasizing the use of Python, Pandas, Scikit-learn, and appropriate hardware specifications. These choices ensure the scalability, performance, and maintainability of the system. The project's contribution to society is significant. By reducing financial losses associated with fraudulent claims, insurers can stabilize premiums for honest policyholders, ensuring fairer rates for everyone. The preservation of trust between insurers and customers is crucial for the long-term

viability of the insurance industry. Successful fraud detection not only protects insurers' bottom lines but also safeguards the interests of policyholders. Additionally, the project enhances resource efficiency by automating fraud detection processes, allowing insurers to allocate resources more effectively . Furthermore, the collaboration between insurance companies and law enforcement agencies can contribute to broader efforts to combat insurance fraud, leading to a safer and more just society. The outcomes of this project are expected to include improved fraud detection accuracy, a reduction in financial losses, and the adaptation to emerging fraud tactics. These outcomes have the potential to reshape the insurance industry's approach to fraud detection, leading to more secure and reliable insurance services for all stakeholders.

REFERENCES

[1] Irum Matloob, Shoab Ahmed Khan, and Habib Ur Rahman. Sequence mining and prediction-based healthcare fraud detection methodology. *IEEE Access*, 8:143256–143273, 2020.

[2] Tessy Badriyah, Lailul Rahmaniah, and Iwan Syarif. Nearest neighbour and statistics method based for detecting fraud in auto insurance. In *2018 International Conference on Applied Engineering (ICAE)*, pages 1–5, 2018.

[3] Archana Singh and Rakesh Kumar. Heart disease prediction using machine learning algorithms. In *2020 International Conference on Electrical and Electronics Engineering (ICE3)*, pages 452–457, 2020.

[4] Bouzgarne Itri, Youssfi Mohamed, Qbadou Mohammed, and Bouattane Omar. Performance comparative study of machine learning algorithms for automobile insurance fraud detection. In *2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)*, pages 1–4, 2019.

[5] Riya Roy and K. Thomas George. Detecting insurance claims fraud using machine learning techniques. In *2017 International Conference on Circuit ,Power and Computing Technologies (ICCPCT)*, pages 1–6, 2017.

[6] Tessy Badriyah, Lailul Rahmaniah, and Iwan Syarif. Nearest neighbour and statistics method based for detecting fraud in auto insurance. In *2018 International Conference on Applied Engineering (ICAE)*, pages 1–5, 2018.

[7] Deepak Kumar Patel and Sharmila Subudhi. Application of extreme learning machine in detecting auto insurance fraud. In *2019 International Conference on Applied Machine Learning (ICAML)*, pages 78–81, 2019.

[8] Abhijeet Urunkar, Amruta Khot, Rashmi Bhat, and Nandinee Mudegol. Fraud detection and analysis for insurance claim using machine learning. In *2022 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*, volume 1, pages 406–411, 2022.

[9] Najmeddine Dhieb, Hakim Ghazzai, Hichem Besbes, and Yehia Massoud. A very deep transfer learning model for vehicle damage detection and localiza-

tion. In *2019 31st International Conference on Microelectronics (ICM)*, pages 158–161, 2019.

[10] Navin Ramesar, Shiva Ramoudith, Nirvan Sharma, and Patrick Hosein. A cost-minimization approach to automobile insurance fraud detection. In *2023 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD)*, pages 1–6, 2023.

[11] Shanthini M. Stacking classifier-based automated insurance fraud detection system. In *2022 IEEE Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, pages 1–6, 2022.

[12] Anisha Singhal, Neha Singhal, Divya, and Kanchan Sharma. Machine learning methods for detecting car insurance fraud: Comparative analysis. In *2023 3rd International Conference on Intelligent Technologies (CONIT)*, pages 1–5, 2023.

[13] Rahul Chauhan, Rohit Negi, and Deepak Kumar Verma. Analysis of machine learning algorithms for insurance fraud detection. In *2023 IEEE 11th Region 10 Humanitarian Technology Conference (R10-HTC)*, pages 649–655, 2023.

[14] Manish Kumar Thukral. Security and efficiency in vehicle insurance: A blockchain-based solution. In *2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*, pages 1129–1136, 2023.

[15] K. Supraja and S.J. Saritha. Robust fuzzy rule based technique to detect frauds in vehicle insurance. In *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, pages 3734–3739, 2017.

[16] Slokashree Padhi and Suvasini Panigrahi. Decision templates based ensemble classifiers for automobile insurance fraud detection. In *2019 Global Conference for Advancement in Technology (GCAT)*, pages 1–5, 2019.

ANNEXURE A

AWARDS/PARTICIPATION IN PROJECT COMPETITION/EXHIBITION

A.1 AMRUT EXPO, ORGANIZED BY AMRUTVAHINI COLLEGE OF ENGINEERING (AVCOE), SANGAMNER

The Amrut Expo, hosted by Amrutvahini College of Engineering (AVCOE) in Sangamner, is an annual event that serves as a platform for students to showcase their innovative projects, engage with industry experts, and foster collaboration between academia and industry. This year's expo took place on January 19th and 20th, 2024, spanning two days of intensive project exhibitions and interactions. Held at the AVCOE campus, the venue provided a conducive environment for participants to present their projects to a diverse audience comprising students, faculty members, industry professionals, and researchers. The heart of the event was the project exhibition, where participants from various educational institutions displayed their projects across a wide range of domains, including engineering, technology, science, and management. Each project demonstrated cutting-edge research, practical applications, and creative solutions to real-world challenges. A panel of judges, consisting of industry experts, academicians, and professionals, meticulously evaluated the projects based on criteria such as innovation, technical merit, feasibility, presentation quality, and societal impact. Participating teams had the opportunity to receive valuable feedback and insights from the judges, enabling them to further refine their projects and enhance their skills. Additionally, the expo facilitated networking and collaboration among participants, allowing them to interact, exchange ideas, and establish connections with peers, mentors, and potential employers. Industry representatives seized the opportunity to scout for talent and explore partnership opportunities with academic institutions. Overall, the Amrut Expo played a pivotal role in fostering a culture of innovation, collaboration, and excellence within the academic and industrial communities

Amrutvahini Sheti & Shikshan Vikas Sanstha's

AMRUTVAHINI COLLEGE OF ENGINEERING

SANGAMNER- 422 608, Dist- Ahmednagar, Maharashtra, India • www.avcoe.org

AMRUT EXPO 2024

Project Exhibition & Competition

19th & 20th January 2024

CERTIFICATE OF EXCELLENCE

This certificate is awarded to

Mr. / Ms. Nikam Kiran Gorakh

of Computer Engineering Department of Amrutvahini College of Engineering in appreciation for active participation & demonstrating his / her project in **Amrut Expo 2024** on 19th & 20th January 2024.

His / Her project secured ..... position.

Co-ordinator | HOD | Dr. R. S. Tajane, Exhibition Co-ordinator | Dr. M. A. Venkatesh, Principal

In Association with

Amrutvahini Sheti & Shikshan Vikas Sanstha's

AMRUTVAHINI COLLEGE OF ENGINEERING

SANGAMNER- 422 608, Dist- Ahmednagar, Maharashtra, India • www.avcoe.org

AMRUT EXPO 2024

Project Exhibition & Competition

19th & 20th January 2024

CERTIFICATE OF EXCELLENCE

This certificate is awarded to

Mr. / Ms. ...Kangane Machindranath Subhash

of .....Computer Engineering.....Department of

Amrutvahini College of Engineering in appreciation for

active participation & demonstrating his / her project in

Amrut Expo 2024 on 19th & 20th January 2024.

His / Her project secured ..... position.

Co-ordinator | HOD | Dr. R. S. Tajane, Exhibition Co-ordinator | Dr. M. A. Venkatesh, Principal

In Association with

Amrutvahini Sheti & Shikshan Vikas Sanstha's

AMRUTVAHINI COLLEGE OF ENGINEERING

SANGAMNER- 422 608, Dist- Ahmednagar, Maharashtra, India • www.avcoe.org

AMRUT EXPO 2024

Project Exhibition & Competition

19th & 20th January 2024

CERTIFICATE OF EXCELLENCE

This certificate is awarded to

Mr. / Ms. Malekar Omm Kailash

of Computer Engineering Department of

Amrutvahini College of Engineering in appreciation for active participation & demonstrating his / her project in **Amrut Expo 2024** on 19th & 20th January 2024.

His / Her project secured ..... position.

Co-ordinator | HOD | Dr. R. S. Tajane, Exhibition Co-ordinator | Dr. M. A. Venkatesh, Principal

In Association with

Amrutvahini Sheti & Shikshan Vikas Sanstha's

AMRUTVAHINI COLLEGE OF ENGINEERING

SANGAMNER- 422 608, Dist- Ahmednagar, Maharashtra, India • www.avcoe.org

4th Time Accredited by NBA | Accredited with NAAC A+ | Affiliated to | Permanently Affiliated to | 'A' Grade Awarded by | TUV SUD Certified ISO 9001:2015 | Accredited by TCS TATA CONSULTANCY SERVICES | Impaneled with

AMRUT EXPO 2024

Project Exhibition & Competition

19th & 20th January 2024

CERTIFICATE OF EXCELLENCE

This certificate is awarded to

Mr. / Ms. Kharde Vishal Lahanu

of Computer Engineering Department of

Amrutvahini College of Engineering in appreciation for active participation & demonstrating his / her project in **Amrut Expo 2024** on 19th & 20th January 2024.

His / Her project secured ..... position.

Co-ordinator | HOD | Dr. R. S. Tajane, Exhibition Co-ordinator | Dr. M. A. Venkatesh, Principal

In Association with

ANNEXURE B

DETAILS OF THE PAPERS PUBLICATION (IF ANY)

Partici[ated in International Conference on Recent Trends and Advancementin Computing Technologies (ICRTACT) organized by Department Of Computer Engineering, AVCOE, Sangamner on 25th and 26th April 2024.

Amrutvahini Sheti & Shikshan Vikas Sanstha's
Amrutvahini College of Engineering, Sangamner
Sangamner SK, Amrutnagar, Sangamner
Department of Computer Engineering
International Conference on Recent Trends and Advancements in Computing Technologies (ICRTACT-2024)
Certificate
of Appreciation

This is to certify that Mr./Miss./Mrs. Nikam kiran G.
has presented a paper entitled Vehicle Insurance fraud claim Detection Using Machine learning. in International Conference on Recent Trends and Advancements in Computing Technologies (ICRTACT-2024) held on 25th and 26th April 2024 at Amrutvahini College of Engineering, Sangamner.

Dr. R. G. Tambe / Dr. D. R. Patil, Co-Coordinator, ICRTACT-2024
Dr. S. A. Thanekar, Coordinator, ICRTACT-2024
Dr. S. K. Sonkar, Convenor, ICRTACT-2024, HoD, Comp. Engg., AVCOE
Dr. M. A. Venkatesh, Principal, AVCOE, Sangamner

Technically Sponsored by :-

Amrutvahini Sheti & Shikshan Vikas Sanstha's
Amrutvahini College of Engineering, Sangamner
Sangamner SK, Amrutnagar, Sangamner
Department of Computer Engineering
International Conference on Recent Trends and Advancements in Computing Technologies (ICRTACT-2024)
Certificate
of Appreciation

This is to certify that Mr./Miss./Mrs. Kangane Machhindranath S.
has presented a paper entitled Vehicle Insurance fraud claim Detection Using Machine learning. in International Conference on Recent Trends and Advancements in Computing Technologies (ICRTACT-2024) held on 25th and 26th April 2024 at Amrutvahini College of Engineering, Sangamner.

Dr. R. G. Tambe / Dr. D. R. Patil, Co-Coordinator, ICRTACT-2024
Dr. S. A. Thanekar, Coordinator, ICRTACT-2024
Dr. S. K. Sonkar, Convenor, ICRTACT-2024, HoD, Comp. Engg., AVCOE
Dr. M. A. Venkatesh, Principal, AVCOE, Sangamner

Technically Sponsored by :-

Amrutvahini Sheti & Shikshan Vikas Sanstha's
Amrutvahini College of Engineering, Sangamner
Sangamner SK, Amrutnagar, Sangamner
Department of Computer Engineering

International Conference on Recent Trends and Advancements in Computing Technologies (ICRTACT-2024)

Certificate
of Appreciation

This is to certify that Mr./Miss./Mrs. Vishal Kharade
has presented a paper entitled Vehicle Insurance fraud Claim Detection Using Machine learning. in International Conference on Recent Trends and Advancements in Computing Technologies (ICRTACT-2024) held on 25th and 26th April 2024 at Amrutvahini College of Engineering, Sangamner.

Dr. R. G. Tambe / Dr. D. R. Patil, Co-Coordinator, ICRTACT-2024
Dr. S. A. Thanekar, Coordinator, ICRTACT-2024
Dr. S. K. Sonkar, Convenor, ICRTACT-2024, HoD, Comp. Engg., AVCOE
Dr. M. A. Venkatesh, Principal, AVCOE, Sangamner

Technically Sponsored by :-

Amrutvahini Sheti & Shikshan Vikas Sanstha's
Amrutvahini College of Engineering, Sangamner
Sangamner SK, Amrutnagar, Sangamner
Department of Computer Engineering

International Conference on Recent Trends and Advancements in Computing Technologies (ICRTACT-2024)

Certificate
of Appreciation

This is to certify that Mr./Miss./Mrs. Malekar Om Kailash
has presented a paper entitled Vehicle Insurance fraud Claim Detection Using Machine learning. in International Conference on Recent Trends and Advancements in Computing Technologies (ICRTACT-2024) held on 25th and 26th April 2024 at Amrutvahini College of Engineering, Sangamner.

Dr. R. G. Tambe / Dr. D. R. Patil, Co-Coordinator, ICRTACT-2024
Dr. S. A. Thanekar, Coordinator, ICRTACT-2024
Dr. S. K. Sonkar, Convenor, ICRTACT-2024, HoD, Comp. Engg., AVCOE
Dr. M. A. Venkatesh, Principal, AVCOE, Sangamner

Technically Sponsored by :-

ANNEXURE C

PLAGIARISM REPORT FOR THIS REPORT

All must attach certificate/report of Plagiarism issued by Urkund Software. Percentage of Similarity should not be more than 30%

ANNEXURE D

ANY OTHER DOCUMENTATION EVIDENCES RELATED TO PROJECT

Itri, B. and Mohamed, Y. and Mohammed, Q. and Omar, B., Performance comparative study of machine learning algorithms for automobile insurance fraud detection, 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS), 2019, doi=10.1109/icds47004.2019.8942277

Summary : these paper gives brief idea about predicting fraud in automobile insurance claims, applying supervised learning to assess algorithm effectiveness. The study concludes that Random Forest outperforms other algorithms and suggests the need for an updated evaluation method for classification models.

Roy, R., George, K. T, Detecting insurance claims fraud using machine learning techniques. 2017 International Confer- ence on Circuit, Power and Computing Technologies (ICCPCT). doi:10.1109/iccpct.2017.8074258

Summary : Above paper Researhcers chooses sample of more than 500 data. And the data divide into training and testing data. We can see that, compare with the algorithms, decision tree and random forest algorithms; have better performance than naïve bayes.

D. K. Patel and S. Subudhi, "Application of Extreme Learning Machine in Detecting Auto Insurance Fraud," 2019 International Conference on Applied Machine Learning (ICAML), Bhubaneswar, India, 2019, pp. 78-81, doi: 10.1109/ICAML48257.2019.00023

Summary : These paper has seed idea conducted nearest neighbor method and interquartile method to detect fraud in car insurance data. There are 3 (three) methods used: distance based, density based and interquartile range for detecting fraud in benchmarking auto insurance datasets. Experiments were also carried out by considering the effect of the feature selection process on the results of accuracy.

Matloob, S. A. Khan and H. U. Rahman,, Sequence Mining and Prediction-Based Healthcare Fraud Detection Methodology, in IEEE Access, vol. 8, pp. 143256-143273, 2020, doi: 10.1109/ACCESS.2020.3013962

Summary : This paper deals with developing a novel method- ology for detect-

ing anomalous claims in auto insurance records by a neural network based Extreme Learning Machine (ELM). Initially, the raw dataset has undergone a preprocessing proce- dure and divided into the training, validation and testing sets. A pool of trained ELM classifiers is then generated by using different combinations of ELM parameters on the train set.

Melih Kirlidog, Cuneyt Asuk, A Fraud Detection Ap- proach with Data Mining in Health Insurance. Procedia - Social and Behavioral Sciences, Volume 62, Pages 989-994. ISSN 1877- 0428. https://doi.org/10.1016/j.sbspro.2021.09.168.

Summary : The structured approach is used for fraud detection. This approach follows the clinical sequence of treatments for patients in the gynecology department using a graph mining algorithm. The expense feature and other features are selected as discriminating features for performing fraud detection.

T. Badriyah, L. Rahmaniah and I. Syarif, "Nearest Neighbour and Statistics Method based for Detecting Fraud in Auto Insurance," 2018 International Conference on Applied Engineering (ICAE), Batam, Indonesia, 2018, pp. 1-5, doi: 10.1109/INCAE.2018.8579155.

Summary : Analyses for shorter time frames such as one year must also be made. These shorter analyses can be useful for hit-and-run approach useful for the small datasets,Fraud detection and prevention can be beneficial for consumers who have to pay to the fraudsters in the form of higher insurance premiums.

A. Urunkar, A. Khot, R. Bhat and N. Mudegol, "Fraud Detection and Analysis for Insurance Claim using Machine Learning," 2022 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES), THIRUVANANTHAPURAM, India, 2022, pp. 406-411, doi: 10.1109/SPICES52834.2022.977407

Summary : feature choice parameter modification area unit explored with an aim of achieving superior prophetic performance with superior accuracy. The assorted machine learning techniques area unit utilized in the development of accuracy of de-

tection in unbalanced samples. As a system, the info are divided into 3 completely different segments. These area unit loosely coaching, testing and validation.

MD Irshad Hussain B, Pramod Kumar K N, Rakesh U B, Sachinraj M R, Tejas S Patil,Vehicle insurance fraud Detection using Machine Learning,Journal of Emerging Technologies and Innovative Research (JETIR), 2023 JETIR July 2023, Volume 10, Issue 7,JETIR2307882

Summary : The paper proposes using machine learning to detect vehicle insurance fraud by leveraging historical claims data. It emphasizes data preprocessing, feature selection, and model training for effective fraud detection. The approach aims to improve accuracy and efficiency in identifying suspicious patterns, contributing to the ongoing challenge of combating fraudulent activities in the insurance industry.