Optical Character Recognition and Natural Language Classification in Insurance: Deep Learning Architectures for End-to-End Claims Processing Automation

Dr. Nkemjika Ezekwembe, Professor of Computer Science, University of Nigeria, Nsukka

1. Introduction to Claims Processing Automation

Claims processing is a time-consuming and costly task within the insurance industry. The traditional method involves human assessment but is now slowly being replaced with automated systems. It is believed that AI-driven models can help in cutting down the costs that an insurance company pays to employ personnel to process claims, as well as reduce the time taken in claim processing, which, in general conditions, takes weeks. AI and machine learning are already being applied to perform claims processing automation. The traditional method of claims processing is manual, with no assistance from machines. The entire process, from scrutinizing and validating the damages to preparing the estimate, generating invoices, and negotiating with the insured, is performed manually. Automating the claims process and generating reports for image-based damage estimation have been gaining attention in recent years. To bridge the gap between the current process and intelligent automation, various developments are being made using technologies like machine learning and deep learning. Valuation of the asset is a complex issue. Thus, handling large-scale damage assessment in automating the claims process is a challenging task. Traditionally, the handling of claims was purely a manual process with minimal assistance from machines. But with the ever-increasing insurance claims, this manual handling of claims becomes expensive, time-consuming, erroneous, and prone to illegal malpractices. Thus, process development and deployment have become so complex that organizations are now involving technological advancement. In this context, this study aims to introduce an AI and machine-learning-based estimation of claim processing automation that has become the need of the hour for the insurance domain. The role of AI and machine learning is transformative. They are impactful in any scenario, and similarly while handling

insurance claims. Whatever the business domain, reducing claim processing time is always a beneficial aspect. This work discusses the utilization of AI and machine learning to design models driven by efficient loss handling and reducing settlement options by valid insurance, apart from benefits by saving man-hours. We present machine learning methods such as Naive Bayes and others for utility estimation of the appropriate system.

1.1. Challenges in Traditional Claims Management

The management of insurance claims is a time-consuming and resource-intensive process due to the large volume of claims to be processed. Due to manual processes and repetitive coordination, the claims management time is quite long, resulting in dissatisfied customers and high operational costs. The process also carries an inherent risk of errors due to manual intervention and biased decision-making. Organizations need to manually process those claims in a more complex collision scenario to investigate and decide on claims. Therefore, it is difficult to standardize the assessment based on experience, knowledge, and guidelines, which will affect claimant satisfaction. So, the need for in-depth claim management technology is emerging. Technologies involving automation are urgently needed in claims processing. An automatic claim management system utilizes AI for processing damage claims.

Insurance claims processing has been manually performed for many years. Manual claims processing poses significant challenges, which include higher operational costs, longer processing times, and greater potential for manual errors, which adds to the cost of the claims and the overhead of manual intervention in tasks. The manual process of gathering claim details, obtaining insured data, analyzing, assessing damaged vehicles, coordinating grievances, and attending the hearing before the welfare body leads to higher customer dissatisfaction. Finally, the report must be submitted to the check before the claim can be settled. This claim process triggers a yielding of time and complexity in rules, deference to agreements, and appraisal providers when engaging in damaged vehicle decisions. Many variables or facets will influence decision-making, and standardizing it alone is burdensome; the information that will help in the defacement value or decision cannot always be obtained. In most cases, the parties in a claim or claimant expect to be phenotypically unique in their situation, so only experts can help extract an inference report.

2. Machine Learning in Claims Processing

Large volumes of highly unstructured claims data pose their fair share of challenges in the processing pipeline. Conventional analysis methodologies often fail to provide broader insights hidden within the data. Machine learning and AI techniques, on the other hand, are more adept at scrutinizing structured and unstructured data, revealing intricate patterns and establishing connections that are monitored and learned every single day. An automated mechanism that uses these technologies for decision-making holds great merit in the claims processing sector. AI-driven models transcend traditional sets of rules and standard analytics processes. Pattern recognition is central to machine learning. It identifies associations and trends based on historical records, which means it can look into larger datasets and identify weak signals that have higher probabilities.

Machine learning models can drive several high-utility applications in the claims processing lifecycle, such as fraud detection and identification, automated damage assessment, filing case predictions, human engagement, estimates, and automation of claims validation. These are redefining elevated, time-sensitive processes that can impact turnaround times and, eventually, customer satisfaction and business revenue. Machine learning use cases in insurance are vast, and many companies are already leveraging machine learning insights to refine their claims processes. Companies are increasingly using machine learning for claims servicing, which helps them detect fraudulent practices beforehand. These real-time analytics-driven predictive insights offer higher precision in underwriting and dynamic pricing strategies. The prevalence of machine learning will see companies extensively influenced by model readings. Insights will become more meaningful and form an essential part of decision-making, complementing the traditional actuarial rate-making segments in the world of insurance.

2.1. Types of Machine Learning Models

This work focuses on methods to automate certain aspects of claims processing. The foundation of the methods used is based on developing machine learning models from historical claim data. In this subsection, we discuss different types of machine learning models that could be useful in the context of claims processing. For each model type, we discuss the specific scenario in claims that this type of model would be used as well as the common algorithms used to develop the model.

Machine learning models can be categorized based on the behavior they are emulating. There are three main categories: • Supervised learning: Involves building a model to predict a target or outcome variable based on historical claimant and claim data. For example, an insurer might develop a model to predict the likelihood of litigation based on historical claims that went to court. Common algorithms include decision trees, random forests, and gradient boosting machines for predictive modeling, as well as logistic regression and support vector machines for binary classification. • Unsupervised learning: Used to find patterns and clusters among claimant and claim data, without a target variable. Common applications include segmentation and predicting future outcomes based on similar historical contexts. Examples of algorithms include k-means clustering and hierarchical clustering. • Reinforcement learning: Involves creating a model to make sequential decisions in a claim based on historical data, ultimately resulting in a reward or benefit defined by the insurer.

Preparing data for machine learning models is important and can consume a significant amount of time and effort as it requires data cleaning, transformation, and feature engineering. The choice of model to develop must drive the data preparation steps, so it's a good idea to understand the nuances of the models we might consider in order to proceed effectively.

3. Data Preparation and Feature Engineering

The widespread use of AI-driven models in various fields, including claims processing automation, has initiated an urgent demand for ruleless model approaches, such as machine learning and data analysis, to efficiently work with large sets of unstructured data. These techniques must be thoroughly understood by healthcare practitioners to make informed decisions about the application of automated solutions in their claims operations. In an AI-driven environment of claims processing automation, both machine learning and data analysis begin with data preparation.

This tedious process of cleaning and preparing data is significant, as a large portion of the overall time is spent exploring and cleaning the data, leaving less time for actual model building. Successful models, which show high prediction accuracy, depend on the availability of large amounts of well-structured data. Data cleaning includes tasks such as transforming and aggregating data abstractions. Multiple approaches for cleaning the data include the split-apply-combine method, the removal and replacement

methods for missing values, and data transformation techniques. When it comes to building a machine learning model, selecting relevant features is a crucial step. It relies on the correct engineering, selection, and construction of predictors that help in accomplishing a fair evaluation of the final model. Users should better identify a range of relevant data sources, as this might reveal features that can be used to predict the primary target variable. Even variables that have little to no association with the target variable can aid in identifying low-quality data. Maintaining inner data dependencies and solving any potential biases is crucial to further ensure that the new data and the target variables are correctly handled.

3.1. Data Sources for Training

Given the vast potential of machine learning models to discover complex patterns, the abundance of structured and unstructured data collected for claims processing is vital for model accuracy. This data can be separated into various sources, including traditional, emerging, and recent data sources. Traditional sources of data can be harvested from claims records, policy records, customer interactions, financial transactions, legal documents, and loss history databases. New sources of data, resulting from technological advancements, include social media, news, video, sensors and wearables, and telematics. Additional sources of data could come from third parties, companies' partners, or participants of the insurance policy, including survey results from policyholders that collect data on downtime or job type.

To date, most organizations focus on traditional internal data, which has led to a lack of diversity in the available data used to train machine learning models. Using diverse sources of training data can enrich the machine learning models. Data is one of the most significant assets of an organization and comes with a range of legislative, privacy, and ethical factors. One of the most significant concerns for an organization is the quality of the underlying data. Traditionally, data used in business has primarily been structured or transactional data. This is because structured data has a defined length and format and is akin to information that is generally stored in a traditional database. Much of this type of data is commonly captured at the point of interaction with customers and can range from a whole host of factors, including loan application date, credit score, marital status, length of employment, and the amount requested to be borrowed. Acquisition, storage, copying, and moving the data is not a simple task. Structured data can result in

non-relational data due to different sources of the data between multiple departments and deeply increases security and privacy concerns. In addition, two of the main challenges of structured data are the fact that it is generally collected by separate systems that are unconnected to each other. Unstructured data, including text files, Word files, PDFs, photos, and video, can also give business leaders a range of insights into what customers may be thinking.

4. Model Development and Training

When developing a model and training it for a supervised learning task like claims processing automation, several steps are involved. First of all, a model has to be selected and may involve feature selection and hyperparameter tuning. For training, labeled training data, a model, and a trainable system are required. After training has finished, the resulting model's performance has to be assessed on validation data. Based on the obtained results, the model is adapted and the whole process is repeated. This iterative 'model-development-validation' circle is repeated until a final model is chosen. After training has finished, we are left with a trained model that can be applied to validation data. During training, the primary focus is to fit the model to training data. If, however, a model is too closely adapted to the training data, this leads to overfitting. Overfit models tend to perform poorly on new data due to having essentially memorized the training data. Conversely, underfit models lack the ability to capture the underlying patterns in the data and are not able to perform well on the testing data.

Our approach for model building and training/validating is iterative and comprises multiple iterations. One cycle of iteration comprises several substeps, which include feature extraction, model selection, and hyperparameter selection. Each iteration begins with the step of training the model and finishes with evaluating the performance of the trained model, i.e., its generalization capacity, on validation data. Afterwards, the model is integrated into the target application, i.e., the claim management system. Regular training on new data is necessary to ensure adequate generalization, as the new data may differ from historical data in several aspects, such as changes in legislation. Tailored solutions for the requirements and sizes of the data can be found in the technical frameworks and tools. Packages and libraries provide practical frameworks and utilities for developing models and conducting evaluations.

4.1. Hyperparameter Tuning

Hyperparameters are critical to the development of machine learning models, as they influence model performance and the accuracy of the outcomes resulting from data. Furthermore, different hyperparameter settings can lead to error-ridden predictions and model instabilities. Hence, tuning the hyperparameters is an essential step towards developing an AI-driven tool for the automation of claims. There are various techniques to tune hyperparameters: grid search, random search, and Bayesian optimization, to name a few. Grid search assesses a grid of predefined hyperparameters and makes a decision based on the selected function. This approach is computationally extensive but is the most straightforward method for identifying hyperparameters in a controlled environment. Random search, on the other hand, is more suited to the practicalities of adjusting hyperparameters, especially in the context of claims automation. It assesses a random set of scores to determine the best-tuned hyperparameters. Bayesian optimization involves the development of a model that predicts the next most probable best hyperparameters using the learned results of those previously attempted. Like random search, this is another weighted approach that can make better-justified evolutions in the hyperparameters. However, Bayesian optimization can be limited when it looks at only a few options. The capacity to assess hyperparameters objectively is also contingent on the selected metrics for evaluation. Other implications of hyperparameters on the stability and predictivity of models and the importance of these considerations in decision-making are also discussed. Given the impact of hyperparameters on model readiness, viable models should have a small range of performance with the best hyperparameters of a potential 10 sets and an optimal performance difference that does not exceed 10%.

5. Integration and Deployment of AI Models in Claims Processing

Despite the benefits that AI systems can deliver, connecting them to the rest of the claims processing ecosystem can be daunting, especially for firms that don’t have a mature risk function. Training all claim handlers and risk managers in the use of the AI model’s subject matter involves far more than just skills training, like teaching them the details of the AI system's logic. Facilitating the link-up of an AI model to an organization, specifically for claims processing, needs to address the data compatibility issue. Many companies still run operations on legacy systems that have been cobbled together over time; adding in a new engine like a parallel auto-detection claims

processing AI does not work. A good way to integrate AI into an existing claims processing system is through the use of proprietary application program interfaces. By linking systems with cloud-based object storage, the call made to acquire outputs from one system is synchronized with a call to the second so that the data can operate in near-real time. Various APIs can be used for ease of system integration. The best companies have their online tools alongside their legacy systems, and they are then persisting everything in a new platform that is all cloud-based, which means that they can scale as soon as they are ready to sunset the legacy systems; the cloud platform is already in place. Technology has one more improvement frontier available: learning from success and failure and providing a feedback loop for the relationship between the AI model and an organization to evolve. That said, without rigorous experimentation and optimization, there is little organizations can do to scale up the system; the model's training data is limited. However, the infrastructure will require a full database migration and considerable customization work, almost as if replacing systems. Feedback loops through advanced technology can learn from outliers or whether an anomaly that the AI model might have overlooked initially resulted in a valid claim. As you learn, you need to go backwards; it helps you improve accuracy. Furthermore, adding a feedback loop will enable dynamic training so that the AI model improves over time. A dynamic training model never stops learning.

6. Future Direction

The previous sections have given insight into the present phase of AI-driven models in automating the claims processing of healthcare providers. From our discussion, it is clear that the systems and models now available are showing the path for future research and development as well as the areas where existing models can be improved. With the emergence of newer trends in AI, such as natural language processing, utilizing the technology for automating processes related to unstructured data is anticipated. With other developing trends in AI, for example, more accurate predictive analytics, alertness, and deductions, it is expected that the already existing AI-driven models in claims auto-adjudication can escalate predictive accuracy, further decreasing the number of claims that would require human intervention. Data is crucially vital in bringing about interoperability services of various AI-driven components. In the medical domain especially, full interoperability, involving accuracy of documentation, is a work in progress involving numerous stakeholders across healthcare and particularly

health AI. The quality and accuracy of data are pivotal in determining the efficiency of updated or newer autonomous models leading to more automated processes. Furthermore, it may even be possible that the proposed components witnessing an independent or integrated update in a real-time environment become AI models to autonomously conduct decision-making to execute and inform on deployment through the same autonomous process. The seamless introduction of AI models and autonomous decisions for deployment will result in a substantial reduction of human intervention. The power of integrating different AI components to aid and make decisions towards the processing of claims will become more predominant. Healthcare organizations, governments, and stakeholders are focusing on updating and promoting regulatory frameworks to oversee the deployment of and secure systems with autonomous decision-making. The burgeoning interest in the ethical use of AI, especially in healthcare, will shape the future deployment and decision-making capacities of AI-driven models. Cutting down the processes on the most minimal touch philosophy, the future direction and deployment of such models will pave the way for potentially engaging them in customer-engaging and personalized service innovations using AI technologies. It can be visualized that claims processes across industries will become more autonomous with the passing of time. All of these emerging trends in AI and models will affect the horizon for deploying AI-driven models in claims automation. Since changes in AI and regulations involving the deployment of models for autonomous decision-making and processes rigorously take time, the opportunity and interest in AI-driven model deployments in automating claims and healthcare are also set until the future trend achieves complete potential. This is a standpoint where adaptiveness and continuous learning enrich and progress new perspectives on future AI trends, technological innovations, and their impact on claims automation.

7. Conclusion

Our work demonstrates how AI-driven models can transform the way claims processing is automated. It shows how traditional claims management is faced with bottlenecks and discusses the data and model-specific moments with our machine learning solution. Three aspects of the development of these AI-driven models have been presented. The basis for the improvement in operational efficiency in claims management lies in data preparation. For the data, considerations for contextual integration suitable for claims use cases have been presented. The presented machine learning models provide state-of-

the-art movement quality. We discussed advanced integration options to bridge gaps between black-box claims models and operational systems, paving the way to full automation. Our experimental results indicate that even simple learning models are able to substantially help claim experts. This allows us to conclude that there is a need for a stronger emphasis on data management and machine learning to accelerate claims, reduce fraud, and improve decision making and loss prevention for the next best action on insurance claims. The presented methods can bridge actual process automation capabilities in the near term and include more advanced black-box modeling at a later stage, when system changes for full automation can be realized. However, continuous investment in sophisticated AI and machine learning tools, and in the skills of the people who can use them, is highly recommended, especially for management. Progressive, competitive digital insurers have long since begun to regard claims management as a driver of customer satisfaction.