National College of Ireland

Next-Generation Compliance Support Tool: Leveraging Machine Learning to Optimize Implementation and Audit Preparedness

MSc Academic Internship
MSc Cybersecurity

Abdul Basit Dalvi
Student ID: 22134697

School of Computing
National College of Ireland

Supervisor: Mr. Vikas Sahni

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Abdul Basit Dalvi
Student ID:	22134697
Programme:	MSCCYB1_JAN23B_O **Year:** 23-2024
Module:	MSc Academic Internship
Supervisor:	Mr. Vikas Sahni
Submission Due Date:	14 Dec, 2023
Project Title:	Next-Generation Compliance Support Tool: Leveraging Machine Learning to Optimize Implementation and Audit Preparedness
Word Count:	7645 **Page Count:** 19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Abdul Basit Dalvi

Date: 14/12/2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Next-Generation Compliance Support Tool: Leveraging Machine Learning to Optimize Implementation and Audit Preparedness

Abdul Basit Dalvi
22134697

Abstract

The research aims to develop a Next-Gen compliance support tool to tackle the observed challenges in auditing processes across diverse organizations. The study addresses two key issues - Firstly, organizations frequently lack precise knowledge of necessary regulatory requirements aligning with their specific industry sector or scope. To mitigate this, the research strives to provide upper management with a customized checklist detailing all crucial actions required for compliance in an exhaustive manner. Secondly, automation and simplification of recurring aspects of audits are also targeted by the research acknowledging identical checklist frameworks throughout different organizations. The tool was designed using Machine learning Decision tree model, to optimize the implementation of compliance measures within organizational frameworks which ensured a proactive approach to regulatory requirements. Additionally, the tool aimed to enhance audit preparedness by providing real-time insights into compliance adherence, identifying potential areas of improvement, and streamlining the audit process through intelligent automation. The research yielded promising results, showcasing the efficacy of the Next-Generation Compliance Support Tool in dynamically adapting to diverse compliance scenarios. While the tool's first accuracy rate was a mere 51%, its positive evaluation signposted opportunities for future tuning and growth. The study offers an easy-to-use practical solution which serves as a bridge between compliance standards and effective implementation.

Keywords: Compliance support tool, machine learning, adaptive compliance roadmaps, compliance management, Decision tree model.

1 Introduction

Multifaceted challenges arose in contemporary business environments due to the ever-evolving nature of standards, and industry-specific regulations. The intricate nature of audit tasks adds to the pile, often leading to strenuous efforts for conducting audits and implementing measures. This situation also poses a hurdle for companies trying to distribute their resources effectively without compromising on security sturdiness. In light of these obstacles, this research paper proposed an innovative idea - a cutting-edge compliance support tool that deployed machine learning algorithms as its backbone framework. Going beyond generic solutions, it crafted adaptive roadmaps towards regulatory compliance tailored exclusively around each organization's distinct needs and corresponding industrial scope. The following statistics underline the necessity of superior tools geared toward improved compliance along with harnessing technology like Machine Learning – further aiding optimization processes concerning both implementations as well as preparedness towards audits on all counts:

- 85% of organizations consider cybersecurity a top compliance priority. (Furlong, 2023)
- 60% of business owners say they struggle with keeping up with compliance and regulations. (Furlong, 2023)
- Over 41% of organizations list updating policies and procedures as a major compliance challenge. (Hawtrey, 2023)

- 61% of compliance functions say high volumes of regulatory change is their biggest challenge. (Hawtrey, 2023)

1.1 Motivation

The motivations driving this project stemmed from a pressing need for a comprehensive compliance framework that can effectively address the limitations of existing tools. The intricate nature of regulatory landscapes demands a sophisticated solution capable of offering nuanced and context-specific recommendations. The urgency arises from the inadequacies of traditional compliance management tools, which often fall short in providing adaptive guidance tailored to the specific requirements of diverse organizations. Leveraging machine learning in compliance management represents a strategic advantage in this context. Machine learning algorithms exhibit the capacity to intelligently analyze vast and diverse datasets, enabling the identification of patterns, correlations, and anomalies.

1.2 Gap in current Literature

The examination of existing compliance management tools reveals a landscape characterized by the continual evolution of regulatory frameworks and the persistent struggle to meet the dynamic needs of organizations. A comprehensive review of these tools underscores their pivotal role in facilitating adherence to regulatory standards, yet it also unveils significant gaps in their adaptability and contextual relevance. Traditional tools often employ rule-based systems that may prove inadequate in navigating the intricacies of rapidly changing compliance requirements. Furthermore, the lack of intelligent, learning-driven capabilities impedes their ability to provide context-specific recommendations tailored to the unique characteristics of diverse industries and organizational structures. The next-generation compliance support tool, by addressing the limitations of existing tools, offers tailored and context-aware compliance recommendations. Industries, irrespective of their regulatory domain, can thus harness the tool's capabilities to streamline their governance processes, enhance audit preparedness, and proactively navigate the intricate landscape of regulatory compliance.

1.3 Research Question and Objective

The central research question guiding this project revolves around the development of a next-generation compliance support tool and its utilization of machine learning algorithms. The primary inquiry was articulated as follows:

"How can a next-generation compliance support tool leverage machine learning algorithms to generate adaptive compliance roadmaps that streamline implementation efforts, enhance audit preparedness, and automatically identify security controls from various regulatory standards, matching an organization's unique compliance requirements?"

In order to confront the real-world difficulties of managing compliance, specific research objectives were devised. The chief objective revolved around constructing a user-centric Next-Generation Compliance Support Tool in which machine learning serves as an integral part for providing bespoke guidance. It further aimed to smoothen out audit operations with regards to compliance and standards, by implementing automated checklists, creating customized methodologies via AI/ML techniques while ensuring complete transparency across all levels of management.

1.4 Methodology

The methodology adopted for the development of the next-generation compliance support tool encompassed several key components, each contributing to the tool's effectiveness and overall success. The first step involved the generation of synthetic datasets. The synthetic datasets were meticulously designed to encapsulate various organization-specific fields and their correlation with relevant regulatory standards, mirroring the complexities of actual compliance scenarios. The subsequent phase involved the implementation of a machine learning model. Leveraging the insights gained from the synthetic datasets, the model was trained to intelligently analyze and interpret organizational inputs. This training process formed the core of the tool's ability to generate adaptive compliance roadmaps, enhancing its capacity to streamline implementation efforts and improve audit preparedness. To enhance user accessibility and practical utility, a Flask web application was developed as an integral component of the methodology. This refined model was seamlessly integrated into the compliance support tool, equipping it with the ability to automatically identify the most relevant and applicable security controls from diverse regulatory standards.

2 Related Work

2.1 Literature review

The paper authored by Michael P. Papazoglou, (Papazoglou, 2011) delved into the critical realm of business processes, particularly those implemented as Service-Oriented Architectures (SOA). Acknowledging their foundational role in organizations, the author emphasizes the profound impact of laws, policies, and industry regulations on these processes. In comparison to research being performed, this paper provides valuable insights into design-time, compliance verification and root-cause analysis. However, the research focus on a next-generation compliance support tool, integrating machine learning for adaptive compliance roadmaps, distinguishes the approach. The gap lies in the need to bridge the limitations of existing tools, incorporating machine learning capabilities to offer context-aware and adaptive solutions. Inspiration were taken from the high-level declarative patterns introduced in this paper and the research proposes the integration of machine learning algorithms. The compliance support tool aims to dynamically generate adaptive compliance roadmaps, providing tailored recommendations aligned with an organization's unique compliance requirements and scope. This solution seeks to address the limitations identified in existing compliance management tools.

Published in May 2012, the article, (Turetken, Elgammal, Van Den Heuvel, & Papazoglou, 2012) authored by Oktay Turetken, Amal Elgammal, and Willem-Jan van de, addresses the pervasive challenge faced by companies in ensuring compliance with dynamic laws, regulations, and standards within a constantly evolving business and compliance environment. The critical analysis delved into addressing the multifaceted challenges organizations encounter in their pursuit of compliance. Through a detailed examination, the authors revealed insights into the potential of this methodology to facilitate the verification and monitoring of processes against established compliance requirements. While the pattern-based approach addresses the challenges of compliance management, the research aims to augment these efforts by integrating intelligent algorithms to provide adaptive and context-aware solutions. Bridging this gap involved synthesizing the strengths of both approaches for a more holistic compliance management solution. Built upon the insights gleaned from the pattern-based approach introduced in this article, the research suggested a solution that incorporates machine learning algorithms. By fusing the structured patterns with intelligent

algorithms, proposed next-generation compliance support tool seeks to enhance the effectiveness of compliance management, offering a more adaptive and context-aware solution.

The paper, (Alattas, et al., 2022) focuses on the concept of an ML-based approach, aiming to train machines to analyze data, identify associations, and develop the ability to learn. The ML-based approach had been scrutinized for its potential to enhance the efficiency of document analysis, presenting an opportunity for organizations to leverage advanced technologies for compliance purposes. The findings underscored the potential benefits of employing ML-based approaches in compliance assessment. The paper contended that ML has the capability to read and analyze documents, extract pertinent information, and assess evidence validity. The exploration of Natural Language Processing (NLP) further enhanced the findings, emphasizing its role in providing computational capabilities related to human language, such as information extraction from texts and language translation. The ongoing research, centered on a tool employing machine learning for adaptive compliance solutions, aims to address the gap by proposing a more holistic framework that goes beyond evidence extraction. Building upon the automated solution proposed in the paper, the research suggests integrating ML-based evidence extraction into a broader next-generation compliance support tool. By combining the strengths of evidence extraction through ML with adaptive compliance features, organizations can achieve a comprehensive solution for compliance management.

Further, the paper (Emett, Eulerich, Lipinski, Prien, & Wood, 2023) explored the practical implementation of ChatGPT in the internal audit processes of Uniper, a large multinational company. The critical analysis delved into the successful initial tests and emphasized the necessity for auditors to meticulously assess the risks and opportunities associated with ChatGPT utilization. The paper underscored the importance of evaluating factors such as model accuracy, reliability, legal and ethical implications, data privacy, security concerns, and the potential impact of biased or inappropriate responses. The findings highlighted the positive outcomes of the initial tests and demonstrated the helpfulness of ChatGPT across various aspects of the internal audit process. Built upon the successful implementation of ChatGPT in internal auditing, the research suggested extending the application to a comprehensive next-generation compliance support tool. This entailed the exploration of how ChatGPT can be adapted to address broader compliance challenges, leveraging its capabilities to offer context-aware and adaptive compliance solutions. The suggested solution aimed to bridge the identified gap, providing a more holistic approach to compliance management beyond the internal audit process.

The paper, (Murakonda & Shokri, 2020) introduced the ML Privacy Meter, a tool developed by the Data Privacy and Trustworthy ML Research Lab at the National University of Singapore. The document outlined the challenges posed by machine learning models in terms of privacy risks to training data. The critical analysis assessed the inherent privacy risks associated with machine learning models, particularly in the context of information leakage through predictions and parameters. The paper emphasized the importance of regulatory compliance and the need for practitioners to analyze, identify, and minimize threats to data privacy. The authors presented the ML Privacy Meter as a tool that guided practitioners through these processes and permitted the deployment of models with improved accuracy while considering utility-privacy trade-offs. The findings showcased the ML Privacy Meter as an effective tool for quantifying the privacy risks of machine learning models to their training data. While the paper offers a valuable contribution to the field of machine learning privacy, a potential gap lies in the exploration of the tool's compatibility and applicability to emerging regulatory frameworks specific to the research domain. Built upon the ML Privacy Meter's success in aiding regulatory compliance, the research suggests conducting an in-

depth analysis of its adaptability to evolving compliance standards. This involved evaluating the tool's effectiveness in addressing the specific privacy challenges identified in our research domain.

Next paper, (Amariles, Troussel, & Hamdani, 2020) addressed the existing information asymmetry between data subjects and processors, posing a threat to the anticipated benefits of privacy regulations like GDPR. The critical analysis assessed the significant information asymmetry issue and its potential impact on the effectiveness of privacy regulations. The critical analysis also evaluated the practical issues, individual tasks, and comments provided by the Privatech project, examining their relevance and effectiveness in advancing privacy solutions. The paper anticipated that the outlined approach, accompanied by practical insights and ongoing development comments, will contribute to advancing solutions and tools for protecting individual privacy and enhancing data protection rights. While the paper provided a comprehensive roadmap and insights, a potential research gap lies in exploring the adaptability of the proposed approach to specific nuances in the research domain being explored. The research aimed to address this gap by conducting a comparative analysis, evaluating how the Privatech project's roadmap aligns with the unique challenges and requirements identified in our research context. Building upon the Privatech project's roadmap, the research suggested conducting a domain-specific evaluation to enhance the roadmap's applicability. The suggested solution aimed to optimize the implementation of automation and machine learning in compliance generation, ensuring its effectiveness in addressing the specific challenges posed by research domain being explored.

The paper, (Bedi, Goyal, & Kumar, 2020) explored the transformative role of artificial intelligence (AI) in risk management and compliance, particularly focusing on its impact on small and medium enterprises (SMEs). The critical analysis evaluated the progress made by businesses in utilizing large-scale data for risk management while emphasizing the inadequacies of conventional computational methods. The authors argued that AI, with its cognitive analysis capabilities, addresses the limitations of traditional approaches by defining risk factors and accommodating dynamic large-scale data. The findings emphasized that AI is becoming a core service across industries, driving strategies for improved customer satisfaction, operational effectiveness, efficiency, and competitiveness. Regulatory authorities are closely monitoring the potential risks and unintended consequences of AI adoption, posing challenges for industries to strike a balance between supporting innovation and ensuring compliance. Built upon the paper's insights, the research suggested a detailed examination of the applicability of AI in the specific context of the research. This involved identifying potential challenges, tailoring AI solutions to address domain-specific requirements, and ensuring that the proposed strategies align with the goals of our research. The suggested solution aims to optimize the implementation of AI in risk management and compliance within the research domain, ensuring its effectiveness and relevance.

This study, (Hamdani, et al., 2021) presented a comprehensive theoretical framework designed for the implementation and monitoring of GDPR compliance within the data supply chain. The study introduced a formal and substantive method to verify GDPR compliance in privacy policies, with potential adaptability to other compliance documents such as Data Protection Impact Assessments (DPIAs) and Records of Processing Activities (ROPAs). The critical analysis was centered on two significant contributions. Firstly, the authors experiment with the automation of formal compliance checking of privacy policies, proposing a system that combines machine learning and rules to detect GDPR-mandated information. Secondly, the study utilized the OPP-115 taxonomy to encode GDPR rules from Articles 13 and 14, evaluating the system on 30 privacy policies. While the study makes notable progress in automating GDPR compliance checking for privacy policies, a notable gap exists in the absence of a comprehensive corpus of data protection documents from the

data supply chain. Built on the study's findings, the suggested solution involved collaborative efforts to create a new corpus of data protection documents from the data supply chain. This initiative will contribute to the development and refinement of compliance checking tasks, ensuring the applicability of the framework to the unique challenges posed by the industry. Additionally, it was proposed that there could be ongoing collaboration with legal and privacy scholars to evolve the GDPR taxonomy, incorporating insights gained from zero-shot predictions and accommodating the diversity of compliance documents in the data supply chain.

2.2 Summary Table

Below table summarizes the entire literature review:

Table 1: Summary table for Related Work

Paper Name	Findings	Gaps
Making Business Processes Compliant to Standards & Regulations	Introduced a declarative language for expressing compliance concerns in business processes. Developed an interactive graphical prototype for compliance requirements.	Lack of comparison with specific tools or technologies in the domain.
Capturing Compliance Requirements: A Pattern-Based Approach	Proposed a pattern-based approach for ensuring compliance with laws, regulations, and standards in business processes. Introduced a toolset for capturing and managing compliance requirements.	Limited discussion on specific patterns and tools used in the approach.
Extract Compliance-Related Evidence Using Machine Learning	Proposed an automated solution using NLP and preprocessing techniques for extracting compliance-related evidence. Highlighted various machine learning models used in different studies.	Limited discussion on the practical implementation challenges of the proposed solution.
Leveraging ChatGPT for Enhancing the Internal Audit Process	Uniper utilized ChatGPT for internal audit tasks, demonstrating efficiency gains. Emphasized the need for careful evaluation of risks and opportunities associated with ChatGPT.	Limited details on the specific audit tasks performed using ChatGPT.
ML Privacy Meter: Aiding Regulatory Compliance by Quantifying the Privacy Risks of Machine Learning	Introduced ML Privacy Meter for quantifying privacy risks in machine learning models. Emphasized the need for assessing and mitigating privacy risks to comply with data protection regulations.	Limited discussion on the specific technical challenges of integrating ML Privacy Meter into different ML models.
Compliance Generation for	Designed a theoretical	Need for a new corpus of

Privacy Documents under GDPR Designed a theoretical framework for GDPR compliance in the data supply chain	framework for GDPR compliance in the data supply chain. Experimented with a system combining machine learning and rules for formal compliance checking of privacy policies.	data protection documents for comprehensive compliance checking.
Basic Structure on Artificial Intelligence: A Revolution in Risk Management and Compliance	Highlighted the role of AI in risk management and compliance. Emphasized the importance of a two-way learning process between AI specialists and business stakeholders.	Lack of specific examples or case studies demonstrating AI's impact on risk management.

2.3 Summary

In conclusion, the literature review has provided a comprehensive overview of various approaches to compliance in business processes. While each work contributes valuable insights, it is evident that there exists a common gap in practical implementation details, comparative analyses, and comprehensive frameworks. Papazoglou's declarative language (2011) introduces an innovative approach, yet lacks specific tool comparisons. Syed Abdullah et al. (2010) emphasize industry expert opinions with a limited geographic focus. Turetken et al.'s pattern-based compliance approach (2012) lacks detailed information on patterns, while an anonymous source discusses compliance in a changing business environment without specific implementation details. Alattas et al.'s automated ML solution (2022) lacks practical implementation insights, and Emett et al.'s showcase of ChatGPT's efficiency (2023) lacks task specifics. Murakonda and Shokri's ML Privacy Meter (2020) lacks discussion on integration challenges, and Amariles et al.'s GDPR compliance framework (2020) lacks a corpus for comprehensive checking. Bedi et al. (2020) highlight AI's role in risk management without specific examples. Collectively, these works underscore the need for future research to bridge these gaps and provide more practical, comparative, and comprehensive insights into compliance implementation in business processes.

3 Research Methodology

The research methodology employed in this study follows a systematic and rigorous approach to ensure the collection, analysis, and interpretation of data are conducted with precision and reliability. The objective of this research was to develop accurate and efficient machine learning models for classifying companies based on specific criteria. The methodology can be broken down into several key stages, each designed to address a specific aspect of the research problem.

3.1 Problem Definition and Scope:

The first step involved a comprehensive literature review and market analysis to identify the key factors that differentiate companies in the given context. This phase defined the scope of the research problem, outlining the variables and parameters to be considered during the study.

3.2 Considerations

- The cybersecurity compliance standard considered for this research and tool implementation was ISO-27001:2013. Out of 114 controls only few controls (i.e. 21 controls) were selected.
- An example scenario, based on real-life industry norms to demonstrate the tool was being considered. Two companies i.e. company A and B are being evaluated through the tool implementation, details of which are given below.
    - Company A is a small manufacturing company that specializes in producing custom-made automotive parts for classic cars. They have a limited online presence, primarily using a simple website for basic information and contact details. They have a total of 20 employees.
    - Company B is a large e-commerce retailer with a significant online presence, selling a wide range of products. They have 500 employees and processes a large volume of customer data for online orders, payment processing, and customer support. However, they do not handle any financial transactions or sensitive financial data directly, and they outsource payment processing to a third-party payment gateway.

3.3 Data Collection and Generation:

To build accurate and diverse machine learning models, a robust dataset was imperative. Data collection involves gathering real-world and synthetic data sources related to ISO 27001 controls. Synthetic data generation techniques are employed to create diverse datasets, incorporating variations in employee range, network configurations, and security policies. The dataset was meticulously curated, ensuring its representativeness and relevance to real-world scenarios.

3.4 Data Preparation and Selection:

Data preparation involves getting the raw information ready for analysis. In this case, data was loaded from a CSV file, and specific columns are chosen based on their importance. For example, details like the number of employees, branches, and various security factors are selected because they significantly impact the model's predictions.

3.5 Data Preprocessing and Encoding:

Data preprocessing is like cleaning and organizing the data to make it suitable for analysis. One important step is converting categories (like types of network topology) into numbers. This conversion, known as one-hot encoding, helps the computer understand these categories. Additionally, the 'Company' values are transformed into numbers using Label Encoding. This step simplifies the data for the model to work with.

3.6 Model Selection and Training:

The subsequent step involved the training of a Decision Tree Classifier using the generated synthetic dataset. The decision tree model, chosen for its interpretability and ability to handle both numerical and categorical features, utilized features such as employee_range, number_of_branches, and types_of_information to predict the target variable—company. Label encoding was applied to categorical features to convert them into numerical representations, facilitating the training process. The dataset was split into training and testing sets to evaluate the model's performance accurately. The Decision Tree Classifier was then trained on the training set, learning patterns and relationships within the data. The resulting model became capable of predicting the company (A or B) based on the input

features provided. To validate the model, a set of example input values was used, and the predictions were compared against the known outcomes. This iterative process of training and validation ensured that the model generalized well to unseen data and could effectively predict the company affiliation based on the input parameters.

3.7 Flask web app:

The Flask web application exemplified the fusion of data science techniques with user interaction, creating a seamless experience for users to assess and predict company status based on input parameters. Flask, a powerful Python web framework, was utilized to showcase the integration of machine learning models into the web development sphere. It demonstrated the intersection of technology and user-centric design. At its core, the Flask application served as an interface connecting users with complex machine learning algorithms. The backend of the application employed Python libraries such as Pandas, XGBoost, and Scikit-Learn to handle data preprocessing, model training, and predictions. The web interface compriseed intuitive forms where users can input specific values related to employee range, number of branches, security policies, risk assessment procedures, and more. This user-friendly design ensured accessibility for individuals with varying technical backgrounds, enhancing the inclusivity of the application.

3.8 Justifications

- ISO 27001:2013 was selected over 2022 due to widespread industry adoption and practical considerations, as transitioning to the latest 2022 version may pose resource challenges and may not be universally implemented across organizations. Moreover, it has very low impact on the ideology and implementation of the research.
- Selection of limited number of controls from ISO-27001:2013 controls list: ISO 27001 has a total of 114 controls, but due to time constraints and to make the demonstration more manageable, a subset of 21 controls was selected. The focus is on demonstrating collaboration of the inclusion and exclusion of controls i.e. machine learning with audit practices.
- Relevant columns used while implementing code: The relevant columns selected in the synthetic dataset are those that directly impact or influence compliance controls. These columns are crucial for making accurate predictions using the machine learning model. The selected columns represent essential features for compliance assessment, contributing to the effectiveness of the model.
- Using synthetic dataset: Using a synthetic dataset allows for a controlled environment, ensuring a focused demonstration without additional complexities. It helps save time and efforts compared to using real-world data, aligning with the goal of presenting a small demo.
- Model used for ML: The choice of utilizing a Decision Tree model in the compliance support tool finds its justification by drawing insights from the literature review. In the study by Hamdani et al. (2021), a combined rule-based and machine learning approach, particularly employing Decision Trees, was proposed for automated GDPR compliance checking. This precedent aligns with our decision, as Decision Trees are known for their interpretability and effectiveness in handling both numerical and categorical features, as emphasized by the work of Amariles et al. (2020). Furthermore, Turetken et al. (2012) discuss a pattern-based approach for capturing compliance requirements, showcasing the flexibility and adaptability required in compliance scenarios, a characteristic inherent in Decision Tree models. The literature reveals that Decision Trees are well-suited for compliance-related tasks due

to their ability to handle rule-based structures, providing an advantage in interpreting and explaining the decision-making process. The approach is consistent with the findings of Amariles et al. (2020), where the transparency of Decision Trees aligns with the need for explainability in compliance scenarios. Therefore, the decision to employ a Decision Tree model in our compliance support tool is not only informed by machine learning principles but also substantiated by its proven utility in addressing compliance challenges, as evidenced by the literature.

3.9 Addressing Challenges and Limitations:

Throughout the research process, several challenges and limitations were encountered. Addressing these challenges involved innovative problem-solving approaches, including data augmentation techniques, ensemble methods, and hybrid model designs. Additionally, the limitations, such as data imbalance and potential biases, were transparently acknowledged. Mitigation strategies were devised to minimize these limitations' impact on the research outcomes, ensuring the results' reliability and applicability. This research methodology represents a meticulous and comprehensive approach to developing advanced machine learning models for company classification. The outcomesprovide valuable insights into the factors shaping business entities and their strategic decision-making processes. As part of future work, continuous model monitoring and updating strategies will be implemented to maintain the models' relevancy in dynamic business environments. Additionally, exploring emerging techniques, such as deep learning architectures and natural language processing, offers promising avenues for extending the research scope and addressing more intricate classification challenges. This research methodology not only contributes significantly to the academic landscape but also holds immense practical implications for industries and organizations. By embracing a multidisciplinary approach and staying abreast of technological advancements, this research sets a benchmark for future studies in the domain of machine learning-based company classification.

4 Design Specification

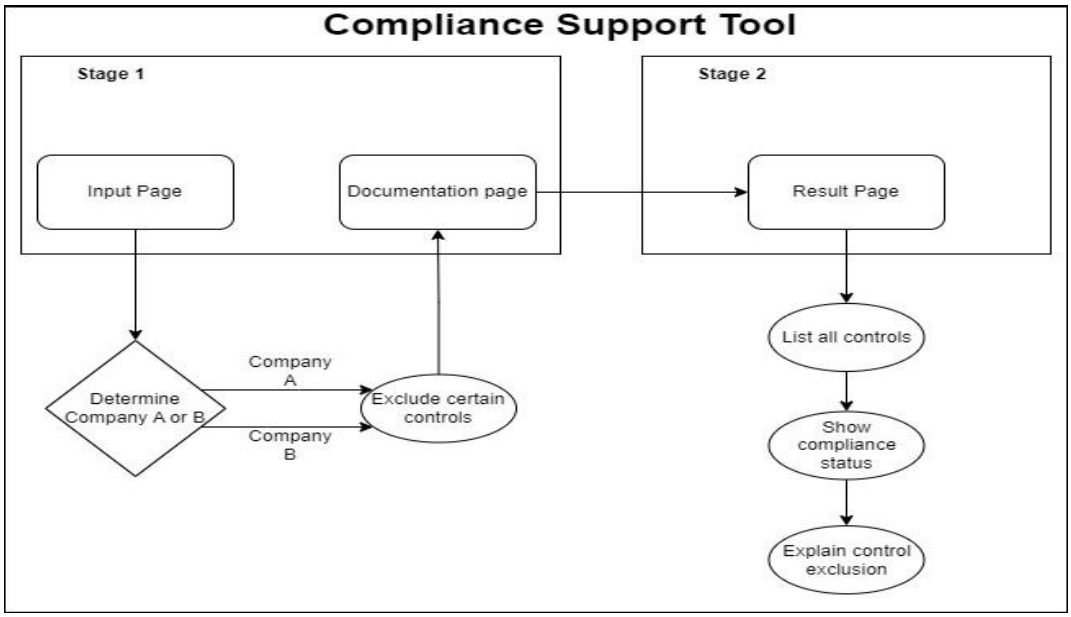

Figure 1: Compliance Support Tool Architecture

The compliance support tool exhibits a well-structured design across three pages:

1. Input Form Page: The implementation design shows that a form where user can enter the details of his company such as company name, no. of employees, branches of company, no. of network devices, no. of workstations, no. of workstations running windows operating system, no. of workstations running on linux operating, information handled by organization, frequency of data transfer, cryptographic controls, event logging mechanism, information security events. This phase takes the primary inputs from the user.
2. Control Exclusion Page: On the basis of the details provided by user the model excludes some of the controls form list of controls(on which it is trained) on the next page . This is based on whether it is company of type A or B. In the next phase the user is asked to fill out the availability and non-availability of required document to fulfill the control. Since the control are already excluded, this makes it easy for the user to fill up all the controls which are required.
3. Documentation and Output Page: After submission of the form the next phase is displaying the result of the status of all controls. Here, this time the tool displays all controls including the controls which are excluded by model. This shows the users the list of controls and to which it is compliant to and not compliant to. The final output is possible due to exclusion of control based on model trained and inputs from user for availability of documentation. Also, proper explanation is provided to the user why the controls are excluded.

Additional Specification:

- Proper validation was taken into consideration to ensure that model was effective enough in selecting the appropriate company type.
- Proper validation was taken into consideration to ensure that model was effective enough in excluding appropriate control.
- It was ensured that tool was scalable enough to handle diverse datasets and flexibility for potential updates or additions to ISO 27001 standards.
- More focus was given on intuitive workflows and user-friendly interactions to enhance the user interface.

5 Implementation

5.1 Dataset Generation

The implementation of the tool began with generating data synthetically and the libraries used for the data generation includes csv, random and faker. The csv library was used to create the csv file of 10000 records of synthetic data which would be generated after the data generation. The random library had been used to provide the random set of values for each row form the given set of values. Faker had been used to generate the fake data. Further, the variables were taken with certain values which were used to generate the random data. After the variables had been set with different sort of values, the data was generated till 10000 rows, writing random data for each row. Post the data had been generated, it is stored in python list and now the data had been written to csv file so that it can further be processed and analyzed for the model generation. Below snippet shows the just a small section of output.csv file which consists of 10000 rows of random data.

```
company,employee_range,number_of_branches,number_of_network_devices,number_of_
workstations,windows_os,linux_os,types_of_information,processes_and_frequency_
of_data_transfers,cryptographic_controls,event_logging_mechanisms,information_
security_events
A,2353,4,15,41,7,34,3,daily,No sensitive financial transactions ,in-house
logging,not required
A,867,4,15,28,1,27,4,no transfer,Good Cryptographic control,not required,not
required
B,1905,8,38,45,34,11,2,daily,Good Cryptographic control,not required,handled
by third-party
```

Figure 2: CSV data

5.2 Model training

In the model generation different libraries of python had been used to create the model such as pandas, scikit-learn. Pandas is the opensource library of python for the manipulation and analysis of the data. After doing the successful import the output.csv had been read using pandas and saved into variable df for further processing. As there were some categorical columns in data, so it was being converted into numerical representation using Label Encoding. After doing the successful conversion from categorical data to numerical representation now the dataframe(df) had been split into two parts - X dataframe and y dataframe. In X dataframe, columns which were independent in nature were being stored. And in y dataframe, there was dependent variable whose value is dependent on independent value. It means the independent variables had direct impact on dependent variable stored in y dataframe. The X dataframe consists of the independent variables such as: employee_range, number_of_branches, number_of_network_devices, number_of_workstations, windows_os,linux_os, types_of_information, processes_and_frequency_of_data_transfers, cryptographic_controls, event_logging_mechanisms, information_security_events. Y dataframe consists of company. As this was the column which the model predicts whether the company was A or B.

After doing the separation, now the data[X,y] was being split into two sets as Train set and Test set. The test size taken was 0.2 which implies 20% of data to be used as test data and 80% of data being used for training of model. The random_state had been set as 42 so that there must be equal representation of each entity of data. After that a Decision Tree model was created and the split data was fed to the model. After the model was trained, it would be tested using the test data i.e. the 20% data which was placed to test the model. After the model is trained and tested using split data, the prediction is validated by using real user input data. The data was passed to model's predict function to get the prediction. The prediction variable provided the predicted company name:

```
Predicted Company: B
```

Figure 3: Tool's prediction

5.3 Flask App Integration

After doing the successful making of model it was to be integrated with the web app so that user can provide the inputs. For this Flask framework had been used which provided the robust approach to make the clear webapps. For this approach, a route was made called as index(1st page, where user input is taken in a form) which takes the values from user for the values on which the model had been trained. The variables were then passed and processed in

the same manner as the model had been built. The data which was given by the user was saved into new_data variable to get the prediction. Again, the categorical data was converted to numerical representation. After that new_data was passed through prediction function so that the prediction can be been made. The company name which was provided as input by user and prediction result are saved in session so that it can be used for controls exclusion. The model was designed in such a way that if Company A is selected, 3 fixed controls are excluded from the control list and if Company B is selected 2 fixed controls are excluded from the control list. Therefore, it means that

- There are two portfolios of companies - company A and company B
- Depending upon the user inputs (due to which scope and scale of the organization was determined), either company A or B is selected.
- And depending on that selection, the controls were excluded. In such manner, the model can be trained on different company portfolios and accuracy and effectiveness of the tool can be improved.

After this the user was redirected to another page for taking the input of controls(2nd page, where user is shown list of controls which are applicable.) The exclusion had been made using the following conditions:

```
{% if session['company_prediction'] == 'A' %}
{% if session['company_prediction'] == 'B' %}
```

Figure 4: Control exclusion

On the second page, user input was taken for availability of document and then redirected to results route(Third page, where output is displayed). Here all the controls are displayed, including the ones which are excluded and the explanation was provided to the user as to why the exclusion had been done.

6 Evaluation

The journey in developing a next-generation compliance support tool had traversed through the intricacies of data generation, model training, and the implementation of a Flask web application. In this section, a comprehensive analysis of the results was obtained from the tool, shedding light on its performance and the consequential implications.

6.1 Evaluation of Model Accuracy

The primary objective for this experiment was to rigorously evaluate the machine learning model's accuracy in categorizing companies A and B based on user input. The utilized dataset for training and testing is characterized by diverse organizational attributes, including employee roles, branch numbers, and types of information. Metrics employed for model evaluation encompass precision, recall, and F1 score, providing a comprehensive performance assessment. Despite achieving an overall accuracy of 0. 51, challenges encountered during the evaluation were acknowledged, informing potential refinements. The results section visually represents the model's performance through insightful charts and graphs, offering a nuanced understanding. The statistical analysis, incorporating precision, recall, and F1 score, accentuates the model's effectiveness. In terms of implications, the discussion revolves around the impact of the model's accuracy on the compliance support tool's overall efficacy. Precise categorization of companies based on user inputs enhances the tool's practical utility, facilitating more informed decision-making in adherence to compliance standards.

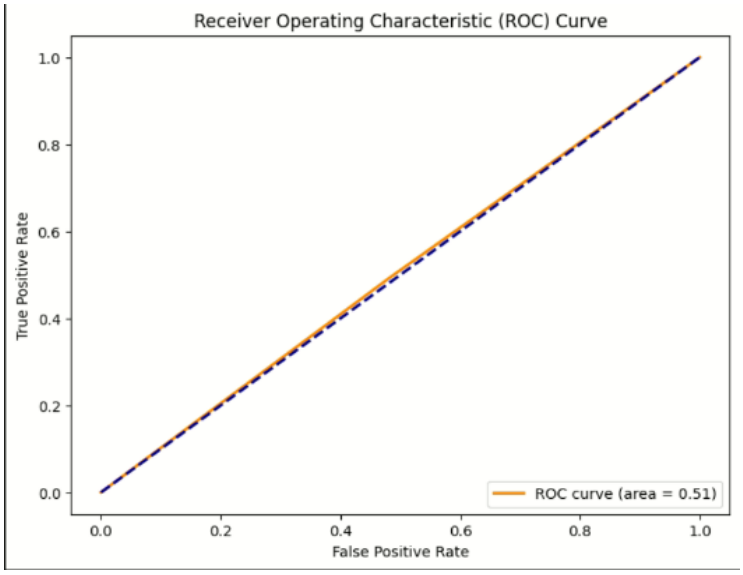

Figure 5: ROC Curve

ROC curve is a straight line, and the area under the curve (AUC) is close to 0.5, it indicates that the model's ability to discriminate between the positive and negative classes is weak as it is the initial phase.

6.2 Control Selection and Documentation Input

Here the objective was two-fold: first, to evaluate the compliance support tool's precision in excluding controls based on the chosen company, and second, to assess the efficacy of the user input process for documenting compliance. Controls were excluded based on the selected company by aligning the specific requirements and characteristics of each company with a predefined set of controls. This process involved tailoring the recommendations to ensure relevance and applicability. The user input process for documenting compliance involved gathering information related to various controls. Users input data based on their company's context, provided details on compliance measures, policies, and procedures. The results included a detailed list of controls deemed applicable for each company. This showcased the tool's ability to precisely exclude controls based on the chosen company. Accurate control selection and effective documentation input were critical contributors to compliance assessment.

6.3 Output Presentation

In Experiment 3, the primary objective was to assess the clarity and user-friendliness of the output presentation generated by the compliance support tool. The output format encompassed the structure and arrangement of information presented to users. This included details on recommended controls, compliance insights, and any additional relevant information. The content was tailored to provide a comprehensive overview of the compliance status. The implications involved a discussion on how a clear and user-friendly output presentation positively influenced user understanding and decision-making. Clear visuals and comprehensible content enhanced the interpretability of compliance recommendations, empowering users to make informed decisions. This user-centric approach contributed to the overall effectiveness and adoption of the compliance support tool in real-world scenarios.

6.4 Discussion

- Evaluating Decision Tree Model Performance: The Decision Tree Classifier demonstrated an accuracy of 51%. While this figure may seem modest, it served as a realistic reflection of the intricate nature of compliance prediction. The Decision Tree's transparency allowed users to comprehend the rationale behind compliance recommendations, fostering trust and user engagement. This opens avenue for further improvements in the tool.
- Striking a Balance between Complexity and Interpretability: The tool's use of machine learning algorithms introduces a tension between model complexity and interpretability. The Decision Tree strikes a balance, offering a transparent decision-making process. However, as the tool evolves, incorporating more sophisticated algorithms for enhanced predictive capabilities necessitates careful consideration of how complexity might impact user understanding.
- Realizing Industry-Specific Adaptations: Acknowledging the diverse landscapes of different industries, the tool aspires to adapt its recommendations to align with sector-specific intricacies. Ongoing efforts include enriching the training dataset with diverse industry inputs, ensuring that the tool evolves into a versatile solution capable of catering to an array of organizational contexts.
- Adapting to Regulatory Dynamism: A critical aspect of the discussion revolves around the tool's agility in the face of dynamic regulatory landscapes. The compliance domain is marked by continuous regulatory updates, demanding a tool that can swiftly align with evolving standards. .
- Addressing Limitations Transparently: A candid discussion about the limitations encountered in the tool's implementation enriches the discourse. From data quality challenges to the trade-offs between interpretability and complexity, each limitation was acknowledged transparently. This not only fosters a culture of openness but also lays the groundwork for targeted improvements. The discussion becomes a compass guiding the tool's evolution towards increased robustness and adaptability.

7 Conclusion and Future Work

The development and evaluation of the compliance support tool have provided valuable insights into its capabilities and areas for future enhancement. The tool, leveraging machine learning and synthetic datasets, demonstrated a 51% accuracy in predicting company compliance profiles. This marks a significant milestone in the integration of artificial intelligence into compliance management, offering organizations a novel approach to streamline and optimize their adherence to standards. The deep content analysis represents a promising avenue for future work. By incorporating advanced content analysis techniques, the tool can evolve to assess the completeness and accuracy of compliance artifacts more comprehensively. This enhancement would address the intricacies of compliance documentation, ensuring that organizations not only meet regulatory requirements but also maintain a robust and effective compliance posture

7.1 Limitations

- The efficacy of any machine learning model is intricately linked to the quality and variability of its training data. Limitations in data quality, such as inaccuracies or biases, can propagate through the model, impacting the accuracy of compliance recommendations.

- Rapid shifts in compliance requirements may introduce a lag as the tool endeavors to align with the latest mandates, necessitating proactive mechanisms to expedite adaptation.
- While the Decision Tree Classifier forms a transparent foundation, the integration of more intricate algorithms might compromise the ease of interpretation.
- Tailoring the tool to match the unique intricacies of each organizational ecosystem remains an ongoing frontier.
- Educating users on the significance of providing accurate and comprehensive inputs becomes crucial to enhance the tool's reliability.
- Handling the escalating complexity of compliance requirements across a burgeoning user base necessitates continuous infrastructure enhancements and optimizations. Striking a balance between real-time responsiveness and scalability becomes pivotal for sustaining the tool's effectiveness.

7.2 Further Improvements

- Deep Content Analysis: Future iterations of the tool should focus on deep content analysis, employing advanced natural language processing (NLP) and machine learning algorithms. This approach can enable the tool to evaluate the semantic context of compliance artifacts, ensuring a nuanced understanding of the information contained within documents. By delving into the depths of content, the tool can provide more accurate assessments of compliance, mitigating the risk of oversights and inaccuracies. The following paper, (Lebanoff & Liu, 2018) can be read for further analysis.
- Multiple Standard/Guidelines Expansion: The tool's potential can be further unlocked by expanding its compatibility to cover an array of standards and guidelines. Beyond the initial selection of ISO-27001, incorporating support for standards like PCI-DSS, GDPR, HIPAA, and others would broaden its applicability. This expansion aligns with the diverse compliance landscape organizations face, offering a comprehensive solution that caters to various regulatory frameworks.
- Best Practices Integration: To enhance its practical utility, the tool should evolve to provide actionable recommendations and best practices. Offering guidance on how organizations can make controls compliant, the tool becomes not just a predictive model but a valuable resource for implementing effective compliance measures. This feature would empower organizations to proactively address compliance challenges, fostering a culture of continuous improvement.
- User Interface Refinement: Improving the tool's user interface and overall user experience is crucial for its widespread adoption. A user-friendly interface with intuitive navigation and clear visualizations will make the tool more accessible to compliance professionals and stakeholders. This, in turn, contributes to the tool's effectiveness in real-world scenarios.
- Integration with Compliance Management Systems: Consideration should be given to integrating the tool with existing compliance management systems. This integration ensures seamless incorporation into organizational workflows, allowing for real-time compliance monitoring and decision-making.

8 References

Alareeni, B. (2019, January 31). A Review of Auditors' GCOs, Statistical Prediction Models and Artificial Intelligence Technology. *International Journal of Business Ethics and Governance, 2,* 19–31. doi:10.51325/ijbeg.v2i1.30

Alattas, H. T., Almassary, F. M., AlMahasheer, N. R., Alammari, R. M., Alswaidan, H. A., Nagy, N. M., . . . Alharthi, S. A. (2022, December 4). Extract Compliance-Related Evidence Using Machine Learning. *2022 14th International Conference on Computational Intelligence and Communication Networks (CICN)* (pp. 537–542). Al-Khobar, Saudi Arabia: IEEE. doi:10.1109/CICN56167.2022.10008324

Amariles, D. R., Troussel, A. C., & Hamdani, R. E. (2020, December 23). Compliance Generation for Privacy Documents under GDPR: A Roadmap for Implementing Automation and Machine Learning. *Compliance Generation for Privacy Documents under GDPR: A Roadmap for Implementing Automation and Machine Learning*. arXiv. Retrieved December 1, 2023, from http://arxiv.org/abs/2012.12718

Bedi, P., Goyal, S. B., & Kumar, J. (2020, December 3). Basic Structure on Artificial Intelligence: A Revolution in Risk Management and Compliance. *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)* (pp. 570–576). Thoothukudi: IEEE. doi:10.1109/ICISS49785.2020.9315986

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., . . . Liang, P. (2022, July 12). On the Opportunities and Risks of Foundation Models. *On the Opportunities and Risks of Foundation Models*. arXiv. Retrieved December 1, 2023, from http://arxiv.org/abs/2108.07258

Bradley, S. (2022). *Resolving conflicts between security best practices and compliance mandates*. (CSO, Editor) Retrieved December 13, 2023, from Resolving conflicts between security best practices and compliance mandates: https://www.csoonline.com/article/573541/resolving-conflicts-between-security-best-practices-and-compliance-mandates.html

Contissa, G., Docter, K., Lagioia, F., Lippi, M., Micklitz, H.-W., Pałka, P., . . . Torroni, P. (2018). Claudette Meets GDPR: Automating the Evaluation of Privacy Policies Using Artificial Intelligence. *SSRN Electronic Journal*. doi:10.2139/ssrn.3208596

Emett, S. A., Eulerich, M., Lipinski, E., Prien, N., & Wood, D. A. (2023). Leveraging ChatGPT for Enhancing the Internal Audit Process – A Real-World Example from a Large Multinational Company. *SSRN Electronic Journal*. doi:10.2139/ssrn.4514238

Eulerich, M., Masli, A., Pickerd, J., & Wood, D. A. (2023, May). The Impact of Audit Technology on Audit Task Outcomes: Evidence for Technology-Based Audit Techniques*. *Contemporary Accounting Research, 40*, 981–1012. doi:10.1111/1911-3846.12847

Furlong, L. (2023, February). *7 Compliance Statistics and What They Mean For You*. Retrieved July 31, 2023, from 7 Compliance Statistics and What They Mean For You: https://thoropass.com/blog/compliance/7-compliance-statistics-and-what-they-mean-for-you/

Governatori, G., & Shek, S. (n.d.). Rule Based Business Process Compliance.

Gu, H., Schreyer, M., Moffitt, K., & Vasarhelyi, M. A. (2023). Artificial Intelligence Co-Piloted Auditing. *SSRN Electronic Journal*. doi:10.2139/ssrn.4444763

Hamdani, R. E., Mustapha, M., Amariles, D. R., Troussel, A., Meeùs, S., & Krasnashchok, K. (2021, June 21). A combined rule-based and machine learning approach for automated GDPR compliance checking. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law* (pp. 40–49). São Paulo Brazil: ACM. doi:10.1145/3462757.3466081

Harkous, H., Fawaz, K., Lebret, R., Schaub, F., Shin, K. G., & Aberer, K. (2018, June 29). Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning. *Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning*. arXiv. Retrieved December 3, 2023, from http://arxiv.org/abs/1802.02561

Hawtrey, D. (2023, January). *Ten policy management and compliance statistics you need to know for 2023*. Retrieved July 31, 2023, from Ten policy management and compliance statistics you need to know for 2023: https://xoralia.com/ten-policy-management-and-compliance-statistics-you-need-to-know-for-2023/

Jan, C.-L. (2021, February 7). Using Deep Learning Algorithms for CPAs' Going Concern Prediction. *Information, 12*, 73. doi:10.3390/info12020073

Lebanoff, L., & Liu, F. (2018, August 28). Automatic Detection of Vague Words and Sentences in Privacy Policies. *Automatic Detection of Vague Words and Sentences in Privacy Policies*. arXiv. Retrieved December 3, 2023, from http://arxiv.org/abs/1808.06219

Ly, L. T., Rinderle-Ma, S., Knuplesch, D., & Dadam, P. (2011). Monitoring Business Process Compliance Using Compliance Rule Graphs. In R. Meersman, T. Dillon, P. Herrero, A. Kumar, M. Reichert, L. Qing, . . . M. Mohania (Eds.), *On the Move to Meaningful Internet Systems: OTM 2011* (Vol. 7044, pp. 82–99). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-25109-2_7

Lynn, T., Mooney, J. G., Rosati, P., & Cummins, M. (Eds.). (2019). Disrupting Finance: FinTech and Strategy in the 21st Century. Cham: Springer International Publishing. doi:10.1007/978-3-030-02330-0

Murakonda, S. K., & Shokri, R. (2020, July 18). ML Privacy Meter: Aiding Regulatory Compliance by Quantifying the Privacy Risks of Machine Learning. *ML Privacy Meter: Aiding Regulatory Compliance by Quantifying the Privacy Risks of Machine Learning*. arXiv. Retrieved December 3, 2023, from http://arxiv.org/abs/2007.09339

Namiri, K., & Stojanovic, N. (2007). Pattern-Based Design and Validation of Business Process Compliance. In R. Meersman, & Z. Tari (Eds.), *On the Move to Meaningful Internet Systems 2007: CoopIS, DOA, ODBASE, GADA, and IS* (Vol. 4803, pp. 59–76). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-540-76848-7_6

Nasr, M., Shokri, R., & Houmansadr, A. (2019, May). Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning. *2019 IEEE Symposium on Security and Privacy (SP)* (pp. 739–753). San Francisco, CA, USA: IEEE. doi:10.1109/SP.2019.00065

Papazoglou, M. P. (2011, August). Making Business Processes Compliant to Standards and Regulations. *2011 IEEE 15th International Enterprise Distributed Object Computing Conference* (pp. 3–13). Helsinki: IEEE. doi:10.1109/EDOC.2011.37

Sadiq, S., Governatori, G., & Namiri, K. (2007). Modeling Control Objectives for Business Process Compliance. In G. Alonso, P. Dadam, & M. Rosemann (Eds.), *Business Process Management* (Vol. 4714, pp. 149–164). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-540-75183-0_12

Shokri, R., & Shmatikov, V. (2015, September). Privacy-preserving deep learning. *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)* (pp. 909–910). Monticello: IEEE. doi:10.1109/ALLERTON.2015.7447103

Song, C., Ristenpart, T., & Shmatikov, V. (2017, September 22). Machine Learning Models that Remember Too Much. *Machine Learning Models that Remember Too Much*. arXiv. Retrieved December 13, 2023, from http://arxiv.org/abs/1709.07886

Strecker, S., Heise, D., & Frank, U. (2011, September). RiskM: A multi-perspective modeling method for IT risk assessment. *Information Systems Frontiers, 13*, 595–611. doi:10.1007/s10796-010-9235-3

Sun, T., & Vasarhelyi, M. A. (2017, June). *Deep Learning and the Future of Auditing*. Retrieved from The CPA Journal: https://www.cpajournal.com/2017/06/19/deep-learning-future-auditing/

Tesfay, W. B., Hofmann, P., Nakamura, T., Kiyomoto, S., & Serna, J. (2018). I Read but Don't Agree: Privacy Policy Benchmarking using Machine Learning and the EU GDPR. *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18* (pp. 163–166). Lyon: ACM Press. doi:10.1145/3184558.3186969

Torre, D., Abualhaija, S., Sabetzadeh, M., Briand, L., Baetens, K., Goes, P., & Forastier, S. (2020, August). An AI-assisted Approach for Checking the Completeness of Privacy Policies Against GDPR. *2020 IEEE 28th International Requirements Engineering Conference (RE)* (pp. 136–146). Zurich: IEEE. doi:10.1109/RE48521.2020.00025

Turetken, O., Elgammal, A., Van Den Heuvel, W.-J., & Papazoglou, M. P. (2012, May). Capturing Compliance Requirements: A Pattern-Based Approach. *IEEE Software, 29*, 28–36. doi:10.1109/MS.2012.45

Zimmeck, S., & Bellovin, S. M. (2011). Privee: An Architecture for Automatically Analyzing Web Privacy Policies.