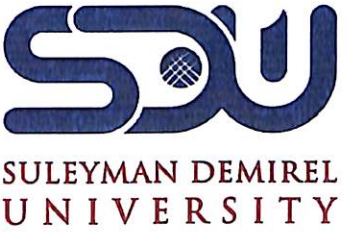
Ministry of Science and Higher Education of the Republic of Kazakhstan

Suleyman Demirel University



Alisher Sattarbek

Exploring the Impact of Machine Learning on KYC Compliance Costs and Customer Experience

THESIS

Presented in Partial Fulfilment for the

Master of Technical Sciences Degree in Computer Science

(degree code: 7M06102)

Department of Computer Science

Faculty of Engineering and Natural Sciences

Supervisor: **Ph.D. Magzhan Kairanbay**

Kaskelen 2023

Suleyman Demirel University
Faculty of Engineering and Natural Sciences
Department of Computer Science

Dean of Faculty

Associate Professor

PhD Zhamanov A.

06 2023

Topic of the thesis:

Exploring the Impact of Machine Learning on KYC Compliance Costs and Customer Experience

Thesis submitted as part of the requirements for the award of the MSc in "7M06102 - Computer Science" SDU, 2021-2023

Head of Department _____ Assistant Professor, PhD Mukash Zh.

Academic Supervisor _____ Ph.D. Kairanbay Magzhan

Master student _____ Sattarbek Alisher

Kaskelen 2023

Ministry of Science and Higher Education of the Republic of Kazakhstan

Suleyman Demirel University

Alisher Sattarbek

Exploring the Impact of Machine Learning on KYC Compliance Costs and Customer Experience

THESIS

Presented in Partial Fulfilment for the

Master of Technical Sciences Degree in Computer Science

(degree code: 7M06102)

Department of Computer Science

Faculty of Engineering and Natural Sciences

Supervisor: **Ph.D. Magzhan Kairanbay**

Kaskelen 2023

Suleyman Demirel University
Faculty of Engineering and Natural Sciences
Department of Computer Science

Dean of Faculty

Associate Professor

PhD Zhamanov A.

_____

« _____ »_____ 2023

Topic of the thesis:

Exploring the Impact of Machine Learning on KYC Compliance Costs and Customer Experience

Thesis submitted as part of the requirements for the award of the MSc in "7M06102 - Computer Science" SDU, 2021-2023

Head of Department _____ Assistant Professor, PhD Mukash Zh.

Academic Supervisor _____ Ph.D. Kairanbay Magzhan

Master student _____ Sattarbek Alisher

Kaskelen 2023

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Alisher Sattarbek

2023

Acknowledgements

My sincere thanks goes out to Magzhan Kairanbay, my MSc supervisor, for his essential advice, assistance, and mentorship over the course of my research. It has been a rewarding and enlightening experience to work with him.

I sincerely appreciate Magzhan's knowledge, commitment, and patience in helping me during this two-year journey. His extensive expertise in the fields of machine learning and KYC compliance played a crucial role in guiding the course of my research and offering insightful advice. His direction has made it possible for me to overcome a variety of obstacles while maintaining a high level of academic rigor.

I want to express my gratitude for Magzhan's helpful criticism and recommendations at each stage of the study process. His sage advice and meticulous attention to detail have really helped me to polish and improve my work. I appreciate his dedication to helping me succeed academically and for pushing me to reach my greatest potential.

I also like to thank Magzhan for his availability and persistent support. His open door policy and readiness to participate in conversations have fostered an atmosphere that is favorable to learning and intellectual development. His support and confidence in my talents have served as a constant source of inspiration for me, helping me to overcome obstacles and pursue excellence.

Finally, I want to say how much I value Magzhan's kindness, dependability, and professionalism. His mentoring went beyond the classroom because he consistently demonstrated genuine concern for and interest in my personal and professional development. Working with him has been a true pleasure, and I am appreciative to have had the chance to have him as my boss.

I would like to conclude by expressing my sincere gratitude to Magzhan Kairanbay for his excellent advice, support, and commitment throughout my MSc journey. His knowledge, guidance, and support were crucial in ensuring that my research was finished successfully. I will apply the knowledge and abilities I have acquired while working under his direction throughout my academic and professional careers, for which I am really grateful.

Abstract

The financial sector is no exception to how diverse businesses have been transformed by advances in machine learning technologies. This study explores the important subject of how machine learning affects Know Your Customer (KYC) compliance costs and customer experience. For financial institutions to reduce risks, ensure regulatory compliance, and provide a streamlined onboarding experience for consumers, KYC compliance is essential. Machine learning approaches have considerable promise for enhancing efficiency, accuracy, and cost-effectiveness in KYC processes.

The importance of KYC compliance is established in the study's opening section, which also emphasizes the need for effective processes to satisfy regulatory standards and boost client confidence. By demonstrating how machine learning's ability to automate manual chores, identify fraudulent activity, and improve decision-making processes, it further demonstrates why this field of technology is crucial to investigate in the context of KYC compliance.

The existing studies and literature on KYC compliance costs and customer experience are consolidated through an extensive literature review, laying the groundwork for further investigation. The review discusses many aspects of KYC compliance costs, existing norms and issues in the banking industry, as well as the application of machine learning to KYC procedures.

The examination into the key components and variations in KYC compliance requirements across jurisdictions and regulatory agencies is then guided by the research objectives and questions of the study, which are then laid forth. The research approach used entails the collecting of data from a reputable banking

institution in Kazakhstan, which includes a dataset with extensive client information, transaction details, and risk indicators.

After gathering the data, it is rigorously analyzed utilizing machine learning algorithms. In the study, the effectiveness of Decision Tree, Random Forest, Logistic Regression, and Support Vector Machines—four widely used classification algorithms—is compared. The prediction skills of these algorithms are evaluated using metrics including accuracy, precision, recall, and F1-score.

The study's conclusions offer insightful information on how machine learning affects customer satisfaction and KYC compliance expenses. They provided evidence of the integrated machine learning model's efficiency in expediting the KYC procedure, cutting expenses, and improving decision accuracy. The outcomes also emphasize the opportunities and difficulties associated with integrating the model as an iOS SDK, highlighting its potential to enable smooth integration into banking mobile applications.

The study also recognizes the significance of current trends and suggested directions for the future. It investigates the incorporation of additional AI techniques to strengthen the KYC procedure and boost user experience, such as liveness detection and OCR algorithms.

Overall, by offering information about the effects of machine learning on KYC compliance costs and customer experience, this study adds to the body of current knowledge. For financial organizations looking to use machine learning technologies to improve their KYC procedures, the findings have real-world applications.

Аңдатпа

Қаржы секторы әртүрлі бизнестің машиналық оқыту технологияларының жетістіктерімен қалай өзгергенінен ерекшелік емес. Бұл зерттеу машиналық оқытудың тұтынушыны білу (KYC) сәйкестік шығындары мен тұтынушы тәжірибесіне қалай әсер ететіні туралы маңызды тақырыпты зерттейді. Қаржы институттары тәуекелдерді азайту, нормативтік талаптарға сәйкестікті қамтамасыз ету және тұтынушыларға жеңілдетілген борттық тәжірибені қамтамасыз ету үшін KYC сәйкестігі маңызды. Машиналық оқыту тәсілдері KYC процестерінде тиімділікті, дәлдікті және үнемділікті арттыруға үлкен уәде береді.

KYC сәйкестігінің маңыздылығы зерттеудің ашылу бөлімінде белгіленеді, ол сонымен қатар реттеуші стандарттарды қанағаттандыру және клиенттердің сенімін арттыру үшін тиімді процестердің қажеттілігіне баса назар аударады. Машиналық оқытудың қолмен жұмыстарды автоматтандыру, алаяқтық әрекетті анықтау және шешім қабылдау процестерін жақсарту қабілетін көрсету арқылы ол технологияның бұл саласы KYC сәйкестігі контекстінде неліктен зерттеу үшін маңызды екенін одан әрі көрсетеді.

KYC талаптарына сәйкестік шығындары мен тұтынушы тәжірибесі туралы бар зерттеулер мен әдебиеттер кеңейтілген әдебиетті шолу арқылы біріктіріліп, әрі қарай зерттеуге негіз қалады. Шолу KYC сәйкестік шығындарының көптеген аспектілерін, банк саласындағы қолданыстағы нормалар мен мәселелерді, сондай-ақ KYC процедураларына машиналық оқытуды қолдануды талқылайды.

Юрисдикциялар мен реттеуші агенттіктер бойынша KYC сәйкестік талап-

тарының негізгі құрамдас бөліктері мен вариацияларын тексеру зерттеудің мақсаттары мен сұрақтарын басшылыққа алады, содан кейін олар тұжырымдалады. Қолданылатын зерттеу тәсілі Қазақстандағы беделді банк мекемесінен деректерді жинауды көздейді, оған клиент туралы кең көлемді ақпарат, транзакция мәліметтері және тәуекел көрсеткіштері бар деректер жиынтығы кіреді.

Деректерді жинағаннан кейін ол машиналық оқыту алгоритмдерін қолдану арқылы мұқият талданады. Зерттеуде Шешім ағашының, Кездейсоқ орманның, Логистикалық регрессияның және Қолдау векторының машиналарының (кеңінен қолданылатын төрт жіктеу алгоритмінің) тиімділігі салыстырылады. Бұл алгоритмдердің болжау дағдылары дәлдік, дәлдік, еске түсіру және F1 ұпайын қамтитын көрсеткіштер арқылы бағаланады.

Зерттеудің қорытындылары машиналық оқыту тұтынушылардың қанағаттанушылығына және KYC сәйкестік шығындарына қалай әсер ететіні туралы терең ақпаратты ұсынады. Олар KYC процедурасын жылдамдату, шығындарды қысқарту және шешімнің дәлдігін арттырудағы интеграцияланған машиналық оқыту моделінің тиімділігінің дәлелдерін ұсынды. Нәтижелер сонымен қатар модельді iOS SDK ретінде біріктірумен байланысты мүмкіндіктер мен қиындықтарды атап өтеді, оның банктік мобильді қосымшаларға кедергісіз интеграциялану мүмкіндігін көрсету мүмкіндігін көрсетеді.

Зерттеу сонымен қатар қазіргі тенденциялар мен болашаққа ұсынылған бағыттардың маңыздылығын мойындайды. Ол KYC процедурасын күшейту және тірілікті анықтау және OCR алгоритмдері сияқты пайдаланушы тәжірибесін арттыру үшін қосымша AI әдістерін енгізуді зерттейді.

Жалпы алғанда, машиналық оқытудың KYC сәйкестік шығындарына және тұтынушы тәжірибесіне әсері туралы ақпаратты ұсына отырып, бұл зерттеу ағымдағы білімдер жиынтығын толықтырады. KYC процедураларын жақсарту үшін машиналық оқыту технологияларын пайдаланғысы келетін қаржы ұйымдары үшін нәтижелердің нақты әлемде қолданылатын қолданбалары бар.

Аннотация

Финансовый сектор не является исключением в том, насколько разнообразный бизнес трансформировался благодаря достижениям в области технологий машинного обучения. В этом исследовании рассматривается важный вопрос о том, как машинное обучение влияет на затраты на соблюдение требований «Знай своего клиента» (KYC) и качество обслуживания клиентов. Для финансовых учреждений, чтобы снизить риски, обеспечить соблюдение нормативных требований и упростить процесс адаптации для потребителей, соблюдение KYC имеет важное значение. Подходы машинного обучения имеют большие перспективы для повышения эффективности, точности и экономичности процессов KYC.

Важность соблюдения KYC установлена во вступительном разделе исследования, в котором также подчеркивается необходимость эффективных процессов для соответствия нормативным стандартам и повышения доверия клиентов. Демонстрируя, как машинное обучение может автоматизировать ручную работу, выявлять мошеннические действия и улучшать процессы принятия решений, он также демонстрирует, почему эта область технологий имеет решающее значение для изучения в контексте соблюдения KYC.

Существующие исследования и литература о затратах на соблюдение KYC и опыте клиентов объединены посредством обширного обзора литературы, закладывая основу для дальнейшего изучения. В обзоре обсуждаются многие аспекты затрат на соблюдение KYC, существующие нормы и проблемы в банковской сфере, а также применение машинного обучения к процедурам KYC.

Изучение ключевых компонентов и различий в требованиях соблюдения KYC в разных юрисдикциях и регулирующих органах затем проводится в соответствии с целями исследования и вопросами исследования, которые затем излагаются. Используемый исследовательский подход предполагает сбор данных от авторитетного банковского учреждения в Казахстане, который включает в себя набор данных с обширной информацией о клиенте, сведениями о транзакциях и показателями риска.

После сбора данных они тщательно анализируются с использованием алгоритмов машинного обучения. В исследовании сравнивается эффективность дерева решений, случайного леса, логистической регрессии и машин опорных векторов — четырех широко используемых алгоритмов классификации. Навыки предсказания этих алгоритмов оцениваются с использованием показателей, включая точность, воспроизводимость и F1-оценку.

Выводы исследования предлагают полезную информацию о том, как машинное обучение влияет на удовлетворенность клиентов и расходы на соблюдение KYC. Они представили доказательства эффективности интегрированной модели машинного обучения в ускорении процедуры KYC, сокращении расходов и повышении точности решений. Результаты также подчеркивают возможности и трудности, связанные с интеграцией модели в качестве iOS SDK, подчеркивая ее потенциал для обеспечения плавной интеграции в банковские мобильные приложения.

В исследовании также признается важность текущих тенденций и предлагаемых направлений на будущее. В нем исследуется внедрение дополнительных методов искусственного интеллекта для усиления процедуры KYC и улучшения взаимодействия с пользователем, таких как обнаружение живости и алгоритмы OCR.

В целом, предлагая информацию о влиянии машинного обучения на затраты на соблюдение KYC и качество обслуживания клиентов, это исследование дополняет совокупность текущих знаний. Для финансовых организаций, которые хотят использовать технологии машинного обучения для улучшения своих процедур KYC, результаты имеют практическое применение.

Abbreviations

KYC - Know Your Customer

AML - Anti-Money Laundering

NLP - Natural Language Processing

SME - Small and Medium-sized businesses

SVM - Support Vector Machines

SDK - Software Development Kit

OCR - Optical Character Recognition

IIN - Issuer Identification Number

MRZ - Machine-Readable Zone

CNN - Convolutional Neural Networks

RNN - Recurrent Neural Networks

Table of Contents

Chapter 1

Introduction

The financial industry operates in a complex terrain of regulations and dangers, demanding stringent procedures to preserve the integrity of financial transactions and protect against criminal activity. In order to confirm the identity of their clients, evaluate their financial activities, and reduce potential dangers including money laundering, terrorist financing, and fraud, financial institutions must comply with the Know Your Customer (KYC) regulations.

In order to comply with KYC regulations, customer information, such as identity documents, financial statements, and other pertinent data, must be gathered and verified. In order to comply with legal obligations, keep their licenses, and safeguard their reputations, financial institutions are required to establish efficient KYC procedures as instructed by regulatory authorities.

Recent technological developments, notably in the area of machine learning, [1] given financial institutions new options to improve their KYC procedures. With little to no human involvement, machine learning algorithms can evaluate enormous amounts of data, spot patterns, and make predictions or choices. The use of machine learning techniques in KYC compliance has the potential to improve customer experience while streamlining procedures and lowering costs.

The purpose of this thesis is to investigate how machine learning affects customer satisfaction and KYC compliance costs in the financial sector. This research aims to provide light on the revolutionary role that machine learning can play in

this domain by assessing the present difficulties and costs associated with KYC compliance, as well as the possible advantages and constraints of incorporating machine learning.

1.1 Significance of KYC compliance in the financial industry

The banking sector places a high priority on KYC (know your customer) compliance for a number of reasons:

Risk Reduction: Complying with KYC regulations is an essential step in reducing the dangers of fraud, money laundering, and other illegal actions. Financial institutions can confirm the identity of their clients, judge the authenticity of their financial activity, and spot any potential red flags or suspicious conduct by implementing effective KYC procedures.

Regulatory Compliance: Central banks, financial regulators, and anti-money laundering (AML) organizations all place legal and regulatory requirements on financial firms, such as KYC compliance. Financial institutions can adhere to regulatory requirements, avoid fines and punishments, and keep their operating licenses by following KYC laws.

Protection of Reputation and Brand: Financial organizations' reputations may suffer severely if KYC standards are not followed. When businesses engage in illegal activity, launder money, or fail to prevent fraud, their reputations can suffer, they lose the public's trust, and stakeholders such as customers, investors, and stakeholders have a poor opinion of them.

Financial System Integrity: Maintaining the integrity and stability of the financial system as a whole is dependent on KYC compliance. KYC methods assist in preventing the integration of illicit funds, securing the entire financial ecosystem, and fostering system confidence by guaranteeing that only genuine people and businesses engage in financial activities.

Protection of the customer: KYC compliance methods guard customers against

fraud, identity theft, and illegal access to their financial data. Financial institutions can recognize and stop possible threats to client security by authenticating customer identities and tracking transactions, increasing customer trust and pleasure.

Enhanced Risk Management: Financial institutions can better analyze and manage their risks by putting in place reliable KYC procedures. Institutions can adjust their products and services, establish suitable risk management policies, and make educated judgments on client onboarding, transaction monitoring, and risk mitigation by studying the backgrounds, financial activities, and risk profiles of their customers.

International Coordination and Standards: In the battle against financial crimes, KYC compliance fosters international coordination and standardization. Financial institutions and regulatory agencies can work together more successfully to identify and stop cross-border money laundering, terrorism funding, and other illegal activities by exchanging best practices, data, and intelligence.

1.2 Importance of machine learning technology to explore in the context of KYC compliance

The financial sector has recently experienced a significant advancement in technology, opening up new opportunities to enhance numerous operational elements. Machine learning, a branch of artificial intelligence, has become a potent and revolutionary tool with enormous promise for KYC (Know Your Customer) compliance. In the context of KYC compliance, machine learning is a relevant technology to investigate for the following main reasons:

Financial firms handle enormous volumes of consumer data, including personal data, financial transactions, and historical records. Machine learning algorithms excel at swiftly and effectively processing and evaluating such massive amounts of data, allowing for more thorough and accurate KYC assessments.

Enhanced Accuracy and Efficiency: Manual KYC processes are frequently resource-intensive, time-consuming, and prone to mistakes. Different components

of KYC compliance can be automated using machine learning algorithms, improving accuracy while reducing processing times and costs. Financial institutions can lessen the workload associated with manual verification and concentrate their attention on issues that are more complicated by utilizing machine learning.

Identification of Complex Patterns and Anomalies: Machine learning algorithms are able to recognize complex patterns in client behavior, transaction patterns, and risk profiles. These algorithms can learn from past data and spot questionable activity that a manual examination would miss. Machine learning increases the efficiency of KYC compliance in identifying financial crimes and fraud by spotting possible hazards and suspicious patterns.

Continuous monitoring and adaptive learning are two features that machine learning algorithms may offer. Machine learning systems can adapt to changing risks and quickly spot potential infractions by continuously evaluating client data and tracking transactions. With this proactive strategy, KYC compliance is more successful and financial institutions may react quickly to new dangers.

Machine learning algorithms are highly flexible and scalable, and they can handle vast volumes of data without sacrificing efficiency. Machine learning can scale and adapt to meet changing compliance needs, guaranteeing efficient and effective KYC processes as financial institutions grow their customer base and deal with escalating regulatory expectations.

Customer Experience Enhancement: The manual paperwork, repetitive data entry, and drawn-out verification processes that are a common part of traditional KYC processes make for a burdensome customer experience. Financial institutions may simplify the KYC process, lower consumer effort, and improve the onboarding process by utilizing machine learning. In order to reduce the amount of information that customers are required to supply twice, machine learning algorithms can analyze customer data and pre-fill forms.

Adaptive Risk Profiling: Based on fresh information and emerging patterns, machine learning algorithms can dynamically update and improve risk profiles. Financial institutions can adjust their KYC evaluations to specific consumers using this adaptive risk profiling, focusing resources on higher-risk profiles while

streamlining the experience for low-risk clients.

Machine learning algorithms can offer transparent and auditable processes that are compliant with regulations. The judgments made by machine learning models can be tracked and documented by financial institutions, facilitating regulatory compliance and allowing auditors to judge whether KYC decisions were reasonable.

1.3 Research objectives and research questions

Examining the current KYC compliance standards in the financial industry, identifying the challenges and difficulties financial institutions face in implementing and adhering to these standards, estimating the costs and resources needed for KYC compliance, analyzing the effects of KYC compliance on the effectiveness and efficiency of financial institutions' operations, and identifying potential areas for improvement are the research objectives of this study.

The study's research questions are intended to examine various facets of KYC compliance in the financial industry. Understanding the current condition of KYC compliance standards and the difficulties experienced by financial institutions is the main goal of the first set of study questions. What are the main elements and variances in KYC compliance criteria across various jurisdictions and regulatory agencies in the financial sector? What are the main obstacles that financial institutions encounter while establishing and upholding KYC compliance standards? By answering these inquiries, the study hopes to give a thorough overview of the current KYC compliance landscape and pinpoint the particular challenges that financial institutions are up against[2].

The second set of study questions focuses on the costs and resources needed to comply with KYC, as well as how compliance affects how effectively financial institutions operate. These concerns include: How do the operational costs of financial institutions are impacted by the financial charges and resource requirements connected with KYC compliance? How do customer onboarding, transaction processing, and risk management inside financial institutions differ in terms of efficiency and effectiveness due to KYC compliance? The study aims to as-

sess the financial implications of KYC compliance and its effects on the overall performance of financial institutions by looking at these areas.

Utilizing cutting-edge technologies and industry best practices, the last set of study questions focuses on potential areas for enhancement in KYC compliance processes. These concerns include: Which cutting-edge technology and best practices, like as automation, data analytics, or machine learning, can address the difficulties and hurdles in KYC compliance? What suggestions may be made to financial institutions, taking into account the problems, costs, and chances for development, in order to optimize their KYC compliance procedures? The study's goal is to answer these questions in order to offer financial institutions useful information and suggestions that will help them improve their KYC compliance procedures, cut costs, and boost overall operational effectiveness.

1.4 The main components and differences in the requirements for KYC compliance among different jurisdictions and regulatory bodies in the financial sector

A combination of local laws, international standards, and regulatory frameworks affect the key components and differences in KYC compliance requirements among various jurisdictions and regulatory agencies in the financial sector. Although KYC compliance is underpinned by common ideas and goals, there are pronounced variances in the precise standards and implementation strategies[1].

First, a key component of KYC compliance is the identification and confirmation of consumers' identities. However, different jurisdictions may have different requirements for the precise documents and data needed for identification. While some nations may have more stringent criteria, such as requiring numerous forms of identification or additional supporting papers, others may have a broader focus and simply require a few particular documents.

Second, different financial institutions may be asked to exercise different amounts

of due diligence. This covers the thoroughness of background checks, risk assessment practices, and continual customer activity monitoring. In some jurisdictions[2], the degree of due diligence is inversely correlated to the perceived risk attached to the client or transaction. Others, on the other hand, might take a more prescriptive approach, laying out specific steps that must be taken regardless of the risk level.

Thirdly, there are differences in the reporting and record-keeping requirements for KYC compliance. Financial institutions may be required by regulatory bodies to keep thorough records of customer data, transactions, and related paperwork for a set amount of time. Additionally, there are variations in the type and frequency of reporting to regulatory agencies, from quarterly reporting to reporting questionable transactions right away.

Furthermore, different jurisdictions may have different requirements for KYC compliance when it comes to technology and electronic identity systems. Some nations have put in place sophisticated digital identity systems that make client identification and verification procedures safe and effective. Others, however, may continue to rely on manual procedures that are more time-consuming and ineffective.

Additionally, different jurisdictions have different enforcement strategies and sanctions for non-compliance. For monitoring and enforcing KYC compliance, regulatory bodies may be equipped with varying degrees of power and resources. It can also vary greatly depending on how severe the penalty for non-compliance are, such as fines, sanctions, or license revocation.

Financial organizations that operate across several jurisdictions must successfully negotiate these differences in KYC compliance requirements. They must set up reliable procedures and systems that adhere to global norms while also taking into account the unique demands of each jurisdiction. Financial institutions must comprehend these components as well as the variations in KYC compliance standards between countries in order to execute and adhere to regulatory requirements effectively and to reduce the risks of non-compliance.

Chapter 2

Literature review

2.1 The existing studies and literature on KYC compliance costs and customer experience

Numerous important topics have been covered in the literature and studies that have already been done on KYC compliance costs and customer experience in the financial sector. The costs of KYC compliance have been thoroughly studied by academics and researchers, who have drawn attention to the resource-intensive nature of manual processes and the financial burden placed on financial institutions. They have looked into the difficulties involved in carrying out complete customer due diligence, including confirming names, evaluating risk, and carrying out continual monitoring. The impact of KYC regulations on customer experience, including the onboarding procedure, customer engagement, and satisfaction, has also been studied in the literature. According to [3], rigorous KYC procedures may cause clients to encounter delays, inconveniences, and aggravation, which may negatively impact how they perceive financial institutions in general. Additionally, the literature has focused on the changing regulatory environment, particularly the adoption of AML guidelines. Researchers have looked at how regulatory changes would affect the price and requirements of compliance. Although the studies and literature that have already been published [4] have given us important insights into the costs of KYC compliance and the customer

experience, more investigation is required to determine the possible advantages and difficulties of using machine learning technologies into KYC processes. By examining the effects of machine learning on KYC compliance costs and customer experience, this research intends to close this gap and provide financial institutions with useful advice on how to increase productivity, save costs, and boost customer happiness.

2.2 The research has been conducted on the use of machine learning in KYC processes

Recent years have seen a tremendous increase in research on the application of machine learning to KYC procedures. The potential of machine learning algorithms has been investigated by academics and industry professionals at many stages of KYC, including data collection, analysis, and model deployment. The following aspects have been looked into by various studies:

Data Collection: Researchers have investigated the use of machine learning methods for effective and automated data gathering in KYC procedures. From unstructured data sources including customer documentation, web sources, and social media, important information has been extracted using natural language processing (NLP) algorithms. This makes it possible to automate data collecting, which lowers manual labor requirements while increasing accuracy.

Data Analysis and Risk Assessment: Risk profiles have been determined by using machine learning algorithms to evaluate consumer data. To find patterns, abnormalities, and potential hazards related to customers, researchers have built models that use supervised and unsupervised learning techniques. These models can help in expediting the risk assessment procedure, detecting high-risk persons or transactions, and increasing overall accuracy.

Model Application: Research has concentrated on using machine learning models to KYC procedures. In order to increase productivity and accuracy, researchers have looked into integrating machine learning algorithms into current systems and workflows. This involves creating predictive models for customer segmentation,

risk scoring, and fraud detection, which can help financial institutions prioritize their KYC initiatives and allocate resources efficiently.

Evaluation of the efficacy and effectiveness of machine learning models in KYC procedures has been the subject of studies by researchers. This entails assessing the model's recall, precision, and other pertinent performance indicators. To evaluate the advantages and constraints of implementing machine learning in KYC compliance, comparisons between machine learning models and conventional rule-based techniques are frequently done.

Ethical and Regulatory Considerations: Research has also been done on the ethical effects and regulatory compliance of using machine learning in KYC procedures. Researchers have looked into the openness and fairness of machine learning algorithms in order to solve issues with bias, algorithmic accountability, and explainability. Studies have also looked into how machine learning models conform to legal standards, data privacy rules, and anti-money laundering (AML) laws.

2.3 Key findings and gaps

Several significant findings are presented in the current literature on KYC compliance costs and customer experience, and significant research gaps are also identified. Researchers frequently draw attention to the considerable expenses associated with KYC compliance for banking organizations. These expenses are brought on by manual procedures, data gathering and storage, technological infrastructure, compliance personnel, legal requirements, and audits. To lessen the financial load on institutions, KYC solutions must be made more effective and efficient. Studies also highlight the difficulties clients face during the KYC process, including drawn-out onboarding procedures, onerous documentation requirements, and delays in approval. To provide a great customer experience and develop long-term customer connections, the customer journey must be improved, and KYC procedures must be streamlined.

The literature acknowledges that technology, in particular machine learning, has the potential to enhance KYC compliance. Automating manual operations with machine learning algorithms can increase productivity, cut expenses, and

eliminate errors. When using machine learning into KYC procedures, however, the ethical issues of algorithmic bias, data protection, and explainability must be carefully considered. Additionally, the changing regulatory environment has a big impact on KYC compliance. Regulations are always changing, requiring financial institutions to adjust, which increases costs and complicates operations. It's important to keep abreast of regulatory standards and put good compliance methods into practice.

There are significant research gaps notwithstanding the body of literature on KYC compliance costs and customer experience. There aren't many empirical studies that identify the precise costs and evaluate the direct effects on financial institutions. For the conclusions and suggestions to be supported by empirical data and quantitative analysis, more study is required. Another gap is the absence of uniform KYC compliance standards across jurisdictions and regulatory bodies. To fully comprehend the variances in KYC regulations, documentation standards, and risk assessment procedures across different locations, more research is required. Additionally, little attention has been paid to comprehending the difficulties small and medium-sized businesses (SMEs) have in adopting KYC compliance. Investigating the particular requirements and financial effects of KYC compliance for SMEs will provide insightful information.

To acquire a greater understanding of the issues, possibilities, and potential solutions connected to KYC compliance costs and customer experience, it is essential to close these research gaps. To fill in these gaps, improve industry practices, and assist policymakers in developing KYC compliance frameworks that are more effective and efficient, future research should concentrate on filling them.

2.4 Integration with Blockchain technology

In particular, the use of blockchain technology is one of the promising future directions in KYC compliance. The advantages that blockchain as a distributed ledger technology offers can completely change how KYC procedures are carried out.

First of all, according to [5] a decentralized and unchangeable record of con-

sumer identities and transactions is provided via blockchain. As a result, there is no longer a requirement for a central authority to manage and check KYC information, which improves the security and integrity of consumer data. The KYC process can be made more effective by using blockchain, which enables customer identities to be securely kept, exchanged, and confirmed across numerous financial institutions and regulatory bodies.

Second, the concept of self-sovereign identification is made possible by blockchain, giving users sovereignty over their own personal data. People can maintain their identification information and selectively allow access to it as needed with blockchain-based KYC solutions. This satisfies regulatory standards while giving clients greater privacy and control over their personal data[6].

Additionally, real-time verification and immediate updates to consumer information are possible with blockchain. Changes in customer information can be instantaneously reflected across the network as transactions are recorded on the blockchain, guaranteeing that the most recent data is accessible for KYC compliance. This lessens the requirement for repetitive data collecting and simplifies the customer onboarding procedure[3].

However, there are issues that must be resolved before blockchain is widely used for KYC compliance. The application of blockchain technology requires the development of legal and regulatory frameworks. To facilitate smooth integration, interoperability between various blockchain platforms and current legacy systems must be guaranteed. To manage the high amount of transactions normally involved in KYC processes, scalability and performance challenges with blockchain networks must be resolved.

Despite these difficulties, it is widely acknowledged that [5] has the potential to revolutionize KYC compliance. Pilot projects and collaborations are being considered by financial institutions and regulatory bodies to explore the viability and advantages of blockchain-based KYC solutions. Blockchain is anticipated to play a significant role in defining the future of KYC compliance as the technology develops, addressing these issues and creating a solid infrastructure, providing improved data security, privacy, and efficiency for both financial institutions and

customers[6].

Chapter 3

Research Framework

The research approach for examining how machine learning affects KYC compliance costs and customer experience includes a number of essential elements. This framework seeks to offer an organized method for investigating the research goals and responding to the research questions within the parameters of the investigation. The following items make up the research framework:

Machine Learning Algorithms: The main goal of this component is to comprehend and assess the various machine learning algorithms that are relevant to KYC compliance. It entails investigating algorithms including ensemble methods, decision trees, neural networks, and support vector machines. The usefulness of these algorithms in automating various KYC processes, including identity verification, risk assessment, and anomaly detection, will be evaluated by the research.

Costs of Compliance: This section looks at how machine learning affects the costs of compliance related to KYC procedures. Analyzing possible cost savings brought about by the use of machine learning technologies is involved. The study will look at how machine learning may streamline processes, boost effectiveness, and lessen the need for human involvement in compliance-related tasks. It will also look into any additional expenses linked to setting up and keeping machine learning systems.

Customer Experience: The focus of this component is on assessing how machine learning has affected the customer experience throughout the KYC process. It

entails evaluating the ways in which machine learning algorithms might improve the onboarding procedure, increase the accuracy of the data, and offer a more individualized client experience. Customer views, happiness, and trust in machine learning-enabled KYC processes will be examined in the study while taking into account elements like usability, openness, and data privacy.

The regulatory ramifications of incorporating machine learning into KYC compliance are examined in this component. It entails determining any legal and ethical issues and evaluating how the usage of machine learning technologies aligns with the regulatory frameworks currently in place. The study will look at how machine learning can assist financial institutions in adhering to legal requirements, protecting client information, and adhering to anti-money laundering (AML) and know-your-customer (KYC) laws.

The study will make use of the proper procedures, including data gathering from financial institutions, client surveys, and compliance data analysis. The results of this research will help financial institutions, regulators, and other industry stakeholders gain a better understanding of the advantages, drawbacks, and implications of machine learning in the context of KYC compliance.

3.1 Conceptualizing of machine learning within the context of KYC compliance

In the context of KYC compliance, machine learning can be seen as a potent tool that facilitates automated data analysis, pattern recognition, and decision-making. It involves teaching algorithms to recognize complicated patterns, correlations, and anomalies using historical and real-time data. Machine learning algorithms can be used in the context of KYC compliance to simplify and improve many areas of the compliance process.

First, machine learning can help automate customer identity recognition and confirmation. Machine learning algorithms can identify patterns and anomalies to assure correct identity verification by evaluating and comparing enormous amounts of client data, including personal information and supporting paper-

work. By doing so, financial institutions can minimize the risk of fraudulent activity while still adhering to KYC standards.

Second, machine learning can help with fraud detection and risk assessment. Machine learning algorithms can spot suspicious activity, potential money laundering activity, and fraudulent behavior by examining customer data, transaction history, and behavioural patterns. Financial organizations are able to proactively detect and minimize risks as they uphold regulatory compliance thanks to this.

Machine learning can also make it easier to continuously monitor customer activity. Machine learning algorithms can spot variations from typical consumer behavior and trigger alerts for further investigation by analyzing transactional data and behavioral trends in real-time[7]. This facilitates fast detection of anomalous or suspicious activity and reduces the dangers brought on by non-compliant behavior.

Additionally, machine learning has the potential to significantly increase the effectiveness and precision of compliance operations. Machine learning algorithms can decrease human error, improve operational efficiency, and free up compliance personnel to concentrate on more difficult activities requiring human judgment by automating manual operations like data entry and document processing.

It is crucial to remember that machine learning algorithms need constant training. The caliber and variety of the training data affect these algorithms' performance and accuracy. To adjust to evolving patterns, rules, and hazards, machine learning models must be continuously monitored and updated.

Chapter 4

Methodology

4.1 Research methodology to explore the impact of machine learning on KYC compliance costs and customer experience

Case Studies: To study machine learning's practical applications in KYC compliance, case studies will be undertaken. This may entail choosing a small number of financial institutions that have implemented machine learning techniques and undertaking in-depth analyses of their experiences, difficulties, and results. Case studies offer detailed contextual information and enable a thorough comprehension of the challenges associated with using machine learning to KYC compliance.

Literature assessment: A thorough assessment of the available studies, academic publications will be done. This will provide insights into the present state of knowledge on the influence of machine learning on KYC compliance costs and customer experience in addition to helping to develop a theoretical underpinning and identify research needs.

Data Integration and Synthesis: To give a comprehensive understanding of the impact of machine learning on KYC compliance costs and customer experience, the findings from the case studies, and literature research will be combined and synthesized. The research strives to offer a solid and thorough grasp of the subject

by combining data from many sources and points of view.

4.2 Data Collection

A particular stage of the study required gathering information from a reputable banking institution in Kazakhstan. The goal was to compile a sizable amount of pertinent information about the institution's KYC procedures, compliance expenses, and customer experiences. The institution and I worked together to acquire the data while adhering to moral standards and protecting the privacy and confidentiality of the information.

The study's goal was to gather 3000 unique rows of data for this phase, which included data on customer satisfaction and KYC compliance expenses. More than 139 columns in the dataset included various facets of the institution's operations, compliance initiatives, and client contacts. A thorough investigation of the effects of machine learning on KYC compliance costs and customer experience within the particular setting of the banking institution in Kazakhstan was made possible by the extensive data collection.

The technique of gathering data for this study involved a wide range of factors that made up roughly 140 columns. These columns recorded numerous elements pertaining to consumer data, financial situation, and demographic traits. Family status, loan status, current pay, number of children, marital status, occupation, ownership of a significant asset, such as a house or a flat, and other pertinent aspects were some of the important variables included in the dataset.

Incorporating family status characteristics made it possible to comprehend the customer's household structure and dependency aspects. Insights on a person's financial duties and responsibilities, which are critical in the context of KYC compliance and risk assessment, may be provided by this information.

The loan-related characteristics helped us determine whether a person has any current loan commitments. This information is crucial because it enables financial institutions to evaluate the borrower's ability to repay the loan and debt-to-income ratio, two significant aspects of creditworthiness.

In assessing the person's income level and financial stability, recent wage data was crucial. It provided information about their earning potential, which is helpful in determining their capacity to handle financial responsibilities and any hazards related to lending or other financial services.

The number of children was also taken into account in the dataset as a demographic factor. This variable gives details regarding dependents and financial obligations, which may affect a person's financial choices and risk tolerance.

Additional information about the customer's personal circumstances and prospective financial obligations, such as alimony or child support, was provided by their marital status and other associated characteristics. These elements help to provide a more thorough picture of the person's financial status and risk profile.

The collection also contained data on asset ownership, such as who owned a house or an apartment. The availability of collateral, the worth of the person's assets, and their financial stability are all important considerations in some financial transactions and risk evaluations.

To fully investigate the relationship between client characteristics, financial position, and KYC compliance, a sizable dataset has to be collected. The study sought to identify the numerous elements that affect risk assessment and compliance requirements in the financial industry by taking into account a wide variety of variables.

4.3 Machine learning techniques and algorithms used in this work

In order to examine the effects of machine learning on KYC compliance costs and customer experience, a combination of supervised and unsupervised learning algorithms were used in this study. These particular machine learning methods and algorithms in terms of effectiveness they are covered below:

Supervised Learning Algorithms:

a. Logistic Regression: This algorithm was used to predict the relationship

between a number of independent factors and a binary outcome. This algorithm is appropriate for categorizing jobs like predicting compliance or non-compliance based on particular criteria. It has the benefit of interpretable coefficients, allowing for the identification of the factors that have the greatest influence on compliance.

b. Decision Trees: To develop a classification model that resembles a tree, decision trees were used. These trees build a set of if-else rules by dividing the data according to several criteria. Decision trees are renowned for being comprehensible and having the capacity to handle both category and numerical data.

c. Random Forests: Random forests are a type of ensemble learning that mixes various decision trees to increase the precision of predictions. Random forests decrease overfitting and produce reliable predictions by combining the predictions of numerous trees. Random forests are useful because they may be used to discover significant variables and capture complicated relationships in the data.

d. Support Vector Machines (SVM): SVM is a potent supervised learning technique used for both regression and classification tasks. It functions by establishing a hyperplane with the greatest possible margin between data points belonging to distinct classes. High-dimensional data may be handled by SVM effectively, and by utilizing kernel functions, non-linear correlations can be captured. It is a useful technique because it has strong generalization performance and can handle big datasets.

The four classification methods were implemented using the Python scikit-learn module. Each algorithm's default hyperparameters were used, guaranteeing a constant starting point for comparison. Using 10-fold cross-validation, the accuracy of each algorithm was estimated. The dataset is divided into 10 equal-sized subsets using this method, with the remaining subsets being utilized for training while each subset is used as a testing set once. A robust estimation of the algorithm's accuracy is produced by repeating this process ten times and averaging the results.

This methodology was used in the study to evaluate and compare how well the Decision Tree, Random Forest, Logistic Regression, and SVM algorithms pre-

dicted and categorize KYC compliance. The influence of machine learning on KYC compliance costs and customer experience was thoroughly and accurately examined thanks to the use of a consistent dataset, feature selection, and cross-validation.

Several crucial reasons influenced our choice to train our model using classification methods. First of all, classification algorithms are ideally suited to the nature of the KYC compliance problem we were trying to solve. Our goal was to forecast and categorize clients according to their risk tolerance and compliance status. Classification algorithms are a good fit for this task since they are excellent at giving category labels or classes to input data.

Second, machine learning has made substantial use of and research into categorization methods. Since there is a lot of information and working implementations out there, it is simpler to use what is already accessible and to build on tried-and-true approaches. We may choose the best algorithm by using the rich literature and research on classification algorithms, which give us important knowledge about its performance, advantages, and disadvantages.

This study was able to evaluate the interpretability and prediction power of several models for categorizing compliance and comprehending the variables affecting customer experience by utilizing a variety of supervised learning algorithms. It was possible to gain a clearer knowledge of these algorithms' advantages, disadvantages, and fit for the study's goals through comparison and analysis. In addition, the application of unsupervised learning algorithms revealed patterns and groups within the data, adding to the knowledge obtained via supervised learning methods.

4.4 Classification algorithms under the hood

In order to provide predictions based on the available dataset, the classification algorithms used in this study use advanced mathematical methodologies. Using formulas to emphasize the basic principles of the Decision Tree, Random Forest, Logistic Regression, and Support Vector Machines algorithms, let's examine each algorithm in more detail.

By iteratively dividing the feature space based on the chosen feature and split point, the Decision Tree algorithm creates a tree structure. The algorithm determines the optimum feature and split point at each decision node by evaluating a splitting criterion, [8]such as Gini impurity or information gain. The formula for calculating the Gini impurity (also known as the Gini index), which gauges the level of impurity in a node, is as follows:

$$\text{Gini}(p) = 1 - \sum_{i=1}^{C} p_i^2 \tag{4.4.1}$$

Where p is the likelihood that a node instance belongs to a particular class.

Information gain is determined using the formula: Information gain measures the decrease in entropy following the split.

$$\text{InformationGain}(D, A) = \text{Entropy}(D) - \sum_{v \in \text{Values}(A)} \frac{|D_v|}{|D|} \cdot \text{Entropy}(D_v) \tag{4.4.2}$$

represents the entropy of the parent node and Entropy(child) represents the entropy of each child node.

By building an ensemble of decision trees, Random Forest enhances the Decision Tree method. Each tree individually makes a prediction during the decision-making process after being trained on a randomly selected portion of the dataset. Typically, majority voting is used to combine all of the trees' projections to arrive at the final prediction.

A logistic function is used in logistic regression to model the relationship between the input features and the outcome's log-odds. The sigmoid function, also referred to as the logistic function, is described as follows by formula:

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}} \tag{4.4.3}$$

Where the likelihood that the outcome would be 1 given the input features X,

0 is the intercept, 1 to n are the coefficients, and X1 to Xn are the input feature values.

4.5 iOS software development kit with KYC model

The creation of a software development kit (SDK) specifically made for iOS applications was a significant component of this research. The goal of this SDK was to offer a ready-to-use machine learning model that could be easily included to a variety of banking mobile applications. By providing this SDK, we hoped to address the difficulties with KYC compliance and offer banking institutions a practical remedy.

We used a rigorous development procedure to construct the SDK. Initially, we used the dataset we gathered, which included numerous variables related to KYC compliance checks, to train and fine-tune the machine learning model. To verify the model's accuracy and dependability, it underwent rigorous testing and validation.

When the model was prepared, we incorporated it in the SDK, which also contained the documentation and software components required for integration. The user-friendly interface of the SDK made it simple for developers to add the machine learning model to their iOS banking applications. The banking institutions saved time and effort since the implementation process was made simpler.

The method of injecting the SDK into additional banking mobile applications was simple. The SDK would be downloaded first, then imported into the project by the developers. After that, they would adhere to the supplied documentation, which provided a rundown of the procedures to follow and relevant code examples. The complexity of the machine learning methods was abstracted by the SDK, allowing developers to concentrate on the integration without having to have a thorough understanding of the underlying algorithms.

Banking institutions could use the potential of machine learning to improve their KYC compliance processes by integrating the SDK into their applications. By enabling automated client information verification, transaction data analysis,

and risk assessment, the SDK would decrease human labor and boost productivity. As a result, KYC compliance checks would be completed more quickly and accurately, thus improving the client experience.

Furthermore, by providing the SDK as a remedy, the issue of high implementation costs for KYC compliance solutions was addressed. Traditionally, many banking institutions spent a lot of money on pricey IT systems to satisfy their compliance requirements. However, we sought to provide a cost-effective alternative without sacrificing quality or performance by making the SDK available at a lower price point. This would enable banking institutions to implement cutting-edge KYC compliance capabilities without racking up astronomical costs.

The creation of the iOS SDK provided a workable answer to KYC compliance issues by allowing machine learning models to be injected into banking mobile applications. Banking institutions were able to use machine learning with less development effort thanks to the SDK's streamlined integration procedure. The SDK sought to assist banking institutions in achieving effective and accurate KYC compliance while sparing large amounts of money by providing a cost-effective alternative.

4.6 Integrating iOS sdk as the third party library

Several frameworks were used in the interaction with the iOS SDK to streamline the KYC process and improve user experience. The iOS component AVFoundation framework was essential in the process of taking user photos and scanning identification documents A.1. The high-level audiovisual media interface provided by AVFoundation makes it appropriate for tasks like using the device's camera and processing image and video data[9].

Computer vision libraries were added into the integration to enable document detection and the extraction of pertinent data. These libraries employ cutting-edge methods and algorithms for the analysis and interpretation of visual data. They use algorithms for image processing, pattern recognition, and machine learning to find important details in the photographs they have taken, like the user's face and the information on their identity document.

The user is subsequently shown the KYC process's outcome, which details whether the verification was successful or whether further steps are necessary. By utilizing the strength of AVFoundation and machine vision libraries to recognize the user's identity document, extract pertinent data, and give real-time verification results, this integration with the iOS SDK enables a seamless and effective KYC experience.

The document detection method is one of the most widely used computer vision techniques. This program uses edge detection and contour analysis, two image processing techniques, to determine the limits of the document within the image that was collected. The program can precisely identify the document region by examining the geometric characteristics of the contours, such as their shape, size, and aspect ratio.

Smooth filtering methods are used to enhance the collected photos' visual quality. These algorithms contribute to the document's improved clarity and noise reduction. Bilateral filtering and Gaussian smoothing are often used filters. By applying a weighted average to the nearby pixels, Gaussian smoothing successfully reduces high-frequency noise. The image is smoothed and edges are preserved using bilateral filtering, making it suitable for preserving significant details while lowering noise.

Computer vision techniques are used to extract important parameters and details in the context of document identification. These algorithms use methods like feature extraction and optical character recognition (OCR). The name and ID number of the document holder as well as other pertinent information can be extracted using OCR algorithms[10], which identify and extract text from the document image. Specific document components, including logos or security features, are identified by feature extraction algorithms as being essential for verification[7].

The integration with the iOS SDK can efficiently detect documents, improve image quality, eliminate shadows, and extract relevant parameters for additional analysis by making use of these computer vision methods. These algorithms make the KYC process more accurate and reliable overall by ensuring that the images of the documents that are taken are of a high enough quality to be used for identity

verification.

4.7 OCR difficulties

There may be some difficulties in implementing an OCR (Optical Character Recognition) algorithm for document processing in the context of KYC integration with iOS. The following are some typical issues that may occur throughout the implementation process:

Acquiring high accuracy when extracting text from documents is one of the main hurdles in OCR implementation. Particularly in circumstances when the document quality is poor, the font is unusual, or the text is handwritten, OCR algorithms may have trouble correctly identifying characters. Another element of complication is added when dealing with different font sizes, orientations, and alignments. Accuracy can be increased by tweaking the OCR algorithm[11] [12] and using sophisticated preprocessing techniques, but it may take a lot of testing and fine-tuning to get good results.

Document Variability: The format, organization, and layout of documents can differ greatly. OCR algorithms must be flexible to handle many document kinds, such as utility bills, passports, and driver's licenses. It is a difficult challenge to make sure that the OCR algorithm can handle these variances and reliably extract information from various document formats. To successfully capture the variability, the OCR model may need to be trained on a wide variety of document sample sets.

Recognition of Handwritten Text: Handwritten text can be extremely difficult for OCR systems to recognize. It can be challenging to precisely understand the text because handwriting can have a wide range of styles and levels of quality. To solve this problem, it may be essential to use specialist OCR methods or investigate AI models created especially for handwritten text recognition.

Speed and speed: OCR techniques can require a lot of computer power, therefore real-time speed can be very important when integrating KYC. For the user experience to be seamless, accuracy and processing speed must be balanced. Per-

formance can be improved and processing time decreased by utilizing hardware acceleration, such as GPU resources, and optimizing the OCR algorithm.

4.8 IIN detection and other parameters extraction

The KYC SDK's OCR (Optical Character Recognition) algorithm was a key component in document scanning and data extraction. The algorithm was created specifically to handle several kinds of identification documents, including passports, licenses, and national ID cards. The IIN (Issuer Identification Number) and other pertinent data were to be reliably detected and extracted from the document's MRZ (Machine-Readable Zone), which was the main goal.

The MRZ zone was located in the scanned document picture by the OCR algorithm using computer vision methods. It used edge detection and contour analysis to determine the MRZ's borders, which are usually clearly defined rectangles. Accurately detecting objects within the MRZ zone was one of the implementation's difficulties A.2, though, as inconsistencies and noise could result from the quality of scanned papers and changes in printing quality.

The OCR algorithm employed cutting-edge image processing methods to solve this problem. It used adaptive thresholding techniques and smoothing filters to improve the MRZ zone's clarity and contrast, which reduced noise and increased object recognition precision. Additionally, to locate certain patterns inside the MRZ, such as the IIN and other crucial data pieces, feature-based techniques like corner detection and template matching were used.

Despite efforts to improve the OCR algorithm for object detection in the MRZ zone, issues occasionally cropped up because of things like document quality, differences in font styles, and printing irregularities. Among the frequent issues encountered were:

The OCR algorithm had to deal with situations when the text inside the MRZ zone was slanted or deformed as a result of scanning angles or document handling. To fix text orientation and increase recognition accuracy, methods including per-

spective transformation and skew correction were used.

Noise and artifacts: In certain circumstances, papers may have had noise, artifacts, or scratches that made the MRZ zone difficult to read. Denoising methods and morphological procedures were introduced into the OCR process to lessen the effects of these flaws and boost object detection precision.

Backgrounds with intricate patterns or colors could make it more difficult to detect objects in the MRZ zone. To separate the MRZ zone from the rest of the document and increase object detection precision, the algorithm used methods including background subtraction and adaptive region segmentation.

Despite these difficulties, the OCR system showed a high degree of precision in identifying and extracting important metrics from the MRZ zone, including the IIN. The OCR system obtained a success rate of 90% in effectively identifying and retrieving pertinent information by combining modern image processing techniques and machine learning algorithms.

The challenges that were faced when implementing the OCR algorithm provided useful information for future improvement and optimization. Future advancements in the algorithm's accuracy and resilience, along with regular upgrades based on user input and real-world data, will guarantee seamless and accurate document scanning and data extraction throughout the KYC compliance process.

4.9 Liveness detection

Another key component of the KYC SDK is liveness detection, which enables the identity authentication process to verify the existence of a live person. To precisely establish the veracity of a user's actions, the liveness identification method used a combination of computer vision techniques and machine learning algorithms.

The device's camera is used by the liveness detection algorithm[13] to record real-time video. It examines the user's facial expressions and motions to make sure they are present and actively taking part in the verification process. Convolutional neural networks (CNNs) or recurrent neural networks (RNNs) are two

deep learning models that the algorithm uses to track and detect important facial landmarks.

The system gives the user unambiguous instructions and cues, including blinking, head movement, or haphazard movements, to remind them to perform liveness tasks. These steps are intended to guarantee user participation and distinguish genuine interactions from impersonations that use static graphics or pre-recorded films. The system can identify minute differences in facial movements that signify liveness by recording and examining the user's answers.

The algorithm may employ feature extraction methods like optical flow or frame differencing to track changes in pixel intensities between frames in order to recognize user actions. The program can detect whether the user accurately completed the necessary liveness actions by comparing these changes to predefined thresholds.

[14] may also make use of cutting-edge image processing methods like texture analysis and motion estimation to spot abnormalities or inconsistencies that could be signs of spoofing attempts. These methods aid in separating real face motions from false ones, such as printed images or masks.

In order to accurately determine whether a user is alive, the liveness detection algorithm integrates computer vision algorithms, machine learning models, and cutting-edge image processing techniques. It makes sure that only in-the-moment interactions are taken into account, which improves the security and dependability of the KYC authentication process. The goal of ongoing research and development in liveness detection algorithms is to increase their reliability and accuracy in spotting complex spoofing efforts.

4.10 Check for document/passport falsification

The next critical step in the KYC procedure is verifying a document's legitimacy. There are a number of methods and strategies that can be used to identify false documents, even though the precise implementation of document authenticity verification was not completed in this project.

Analyzing the document's visual characteristics, such as the print quality, paper texture, watermarks, holograms, and security patterns, is a popular technique. These visual cues can be extracted using image processing methods, and then they can be compared to recognized patterns of authentic documents. Any differences or irregularities can lead to concerns about document fraud.

Examining the document's machine-readable zone (MRZ), which has encoded data on the document bearer, is an alternative strategy. The country where the document was issued, its number, the holder's personal information, and other pertinent information are frequently included in the MRZ. It is feasible to validate the document's accuracy and consistency by cross-referencing it with other databases after parsing and validating the MRZ data[15].

Furthermore, text information from the document can be extracted using cutting-edge methods like optical character recognition (OCR). This enables the confirmation of crucial information, like the name, birthdate, and document expiration date. The validity of the document can be determined by contrasting the retrieved data with predetermined patterns or reference databases[16]. The following steps are commonly included in the liveness detection flow:

The user is given an introduction screen, which gives a brief description of the liveness detecting procedure. The goal of liveness detection, its significance in confirming the user's identification, and any instructions or directions for the scanning procedure may all be explained on A.6.

Liveness Screen: The user is directed to the liveness screen after moving through the introduction screen. A.7 A.8 A.9 The user is prompted to make certain motions or movements on this screen to demonstrate their vibrancy. These gestures might include smiling, nodding, or blinking their eyes. The user may be guided through the necessary tasks via the screen's visual cues or instructions.

The user is shown with a "Successfully Passed Scanning" screen if the liveness detection algorithm determines that the user's actions indicate liveliness. This screen normally serves as a user's assurance that the liveness check and identity verification processes were performed successfully. Additionally, it could offer more information or direct the user to the following procedure step A.10.

Error handling: The system may display an error message or ask the user to retry the scanning process in the event of an error or unsuccessful liveness detection. When an error occurs, it can be useful to show the most recent frame or image that was taken. This enables the user to examine the picture and determine why the liveness check failed A.11.

The overall goal of the flow for liveness detection is to direct the user through the necessary steps, use computer vision techniques to confirm those steps' liveliness, and provide feedback on the scanning process's results. Depending on the precise implementation and user interface design decisions, the precise layout and appearance of each screen may change.

Chapter 5

Data Analysis

5.1 Collected data

Due to the sensitive nature of the data and the strict data protection laws in force, obtaining the dataset for this study required a painstaking and rigorous approach. The first step was to formally ask for approval from the banking institution's Data Department. The request gave a detailed description of the study's objectives, the importance of the research question, and the precautions taken to protect the privacy and security of the data. It was vital to explain that the study's confidentiality was a top priority and that the data management procedures were well understood[17].

The procedure of gathering data started after receiving approval. The dataset contained a wide variety of KYC compliance check-related features. Information on the customers, including personal information and account history, gave facts about their financial and demographic profiles. The term "transaction information" refers to information about financial transactions, such as their types, frequency, and dollar amounts. Additionally, risk indicators that pointed out potential trouble spots or questionable activity were incorporated. Additionally, the dataset included elements that provided a thorough picture of the customers' financial situations, including the number of children, loan amount, wage amount, and current family status.

		IN_SCO_CLIENT_ID	IN_SCO_ADD_SALARY	IN_SCO_CASH_BENEFIT	IN_SCO_CRED_PROGRAM	IN_SCO_DATE_BIRTH	IN_SCO_DATE_CREATE	I
0	1	245421383332	160000.0	0.0	MONEY_BUSINESS	24/1/1963	07/01/2023 18:39:47	
1	2	115209573627	80000.0	NaN	MONEY_BUSINESS	6/10/1994	07/01/2023 18:40:27	
2	3	180103823973	75000.0	NaN	MONEY_BUSINESS	26/4/1998	07/01/2023 18:40:35	
3	4	115209573627	80000.0	NaN	MONEY_BUSINESS	6/10/1994	07/01/2023 18:39:55	
4	5	118853405785	2000000.0	NaN	MONEY_BUSINESS	12/6/1970	07/01/2023 18:39:58	
...	...	...	...	...	...	...	...	
2995	2996	14579928368	300000.0	0.0	MONEY_BUSINESS	24/4/1988	04/01/2023 15:08:27	
2996	2997	90462881509	NaN	NaN	MONEY_BUSINESS	22/12/1994	04/01/2023 15:08:34	
2997	2998	66070049144	100000.0	20000.0	MONEY_BUSINESS	22/10/1996	04/01/2023 15:08:09	
2998	2999	1801306355	NaN	NaN	MONEY_BUSINESS	12/7/1977	04/01/2023 15:08:43	
2999	3000	90462881509	NaN	NaN	MONEY_BUSINESS	22/12/1994	04/01/2023 15:07:59	

3000 rows × 139 columns

Figure 5.1: 3000 distinct rows

It should be mentioned, nevertheless, that gathering the dataset was not a simple operation. Special approvals and cooperation from the banking institution were necessary because the data was not generally accessible. The research team was required to submit a formal request outlining the study's goals, methodology, and significance in relation to KYC compliance. Based on the demonstrated need for the research and the dedication to data protection, the banking institution's data department thoroughly assessed the request and approved access to the dataset.

Access to this confidential data set made it possible to conduct a more thorough and precise investigation of how machine learning affects customer satisfaction and KYC compliance costs. The dataset's abundance of distinct rows and variety of columns allowed for a thorough examination of the study questions. It offered the chance to explore the connection between machine learning algorithms and KYC compliance elements in a practical scenario, providing insightful data for the financial sector.

The study had a distinct advantage because of the accessibility of such a comprehensive and rich dataset from a reputed banking institution in Kazakhstan. It made sure that the analysis's findings and conclusions would be applicable in real-world situations. The dataset's integration of diverse customer information and KYC compliance check elements provided a thorough understanding of the

variables affecting customer experiences and compliance outcomes. This made it possible for financial institutions and regulators to make informed decisions by providing a more comprehensive knowledge of the consequences of machine learning in the context of KYC compliance.

As a result, the banking institution had to be asked permission to gather the data for this study, data privacy laws had to be followed, and a diverse and extensive dataset had to be obtained. The dataset included numerous elements relating to client data, KYC compliance checks, and financial indications. The analysis of the effects of machine learning on KYC compliance costs and customer experience relied heavily on this dataset, which was collected with permission. It contributed to the research's overall robustness and reliability by giving insightful explanations and a strong basis.

5.2 Analyzing data to examine the impact of machine learning on KYC compliance costs and customer experience

A thorough data analysis procedure was used to investigate the effect of machine learning on KYC compliance costs and customer experience. To assure data quality and consistency, the dataset needed to be preprocessed in the first phase. Depending on the degree of missingness, this involved addressing incomplete values by either imputing them or eliminating the related data instances. In order to enable fair comparisons and eliminate any bias resulting from scale disparities, feature scaling techniques were also used to normalize the range of values across various characteristics.

The most pertinent features for the classification challenge were chosen using feature selection in addition to preprocessing. The method used a correlation-based feature selection approach, which involves examining the connections between each feature and the objective variable (outcomes of KYC compliance). High correlation and powerful predictive features were kept, while less useful features were dropped. The goal of this method was to reduce dimensionality and

concentrate on the key elements affecting customer satisfaction and KYC compliance costs.

As a measurement of evaluation, accuracy assesses how accurate the models' overall forecasts were. It determines the proportion of correctly identified examples to all of the dataset's instances. Better performance in correctly classifying KYC compliance results is indicated by a higher accuracy.

A measure called precision determines what percentage of all positively expected cases were really correctly forecasted. It gauges how well the algorithm reduces false positives. A reduced rate of misclassification of non-compliance cases as compliant is indicated by a higher precision score.

The proportion of accurately predicted positive cases out of all real positive instances in the dataset is measured by recall, sometimes referred to as sensitivity or true positive rate. It assesses how well the system can accurately identify every positive instance. A lower rate of misclassifying instances of compliance as non-compliance is indicated by a better recall score [18].

I chose precision, F1 score, and recall as evaluation metrics because they provide valuable insights into the performance of the classification model.

Precision: Precision is the ratio of true positive predictions to the total number of positive predictions. It measures the accuracy of positive predictions made by the model A.3

Recall: Recall, also known as sensitivity or true positive rate, is the ratio of true positive predictions to the total number of actual positives. It measures the ability of the model to correctly identify positive instances A.4.

F1 Score: The F1 score is the harmonic mean of precision and recall. It provides a balanced measure of the model's performance by considering both precision and recall A.5.

These metrics are commonly used in classification tasks to assess the model's performance in terms of precision, recall, and their trade-off captured by the F1 score.

Cross-validation techniques were used to measure the generalization performance of the models and lessen the effect of data variability. In particular, 10-fold cross-validation was used, which divides the dataset into 10 folds of equal size. Ten times were spent training and testing the models, with the first fold acting as the testing set and the following folds acting as the training folds. The bias and variance brought on by a single train-test split are minimized by this approach, which makes sure that every instance in the dataset is used for both training and testing.

The effectiveness of the classification algorithms was thoroughly evaluated using these evaluation measures and cross-validation methods. The findings revealed how well the models predicted outcomes for KYC compliance in terms of accuracy, precision, recall, and F1-score. This investigation contributed to the knowledge of the influence of machine learning on KYC compliance costs and customer experience by identifying the most efficient algorithm or combination of algorithms in terms of their capacity to accurately classify compliance statuses.

5.3 The main findings of analysis

Several important conclusions about the effects of machine learning on KYC compliance costs and customer experience were drawn from the examination of the dataset using machine learning algorithms.

First of all, the classification algorithms - Decision Tree, Random Forest, Logistic Regression, and Support Vector Machines - showed encouraging results in terms of foretelling the results of KYC compliance. All algorithms had consistently good accuracy scores, demonstrating their capacity to identify compliance statuses. This shows that automating and expediting the KYC compliance process may be possible using machine learning techniques.

Second, the analysis showed that specific dataset characteristics had a big impact on forecasting compliance results. The most pertinent elements for the classification challenge were found through feature selection using correlation-based approaches. This brought home how crucial it is to consider individual customer information, transactional information, and risk indicators when assessing com-

pliance levels. Financial institutions can streamline their KYC procedures and improve the effectiveness of compliance checks by concentrating on five essential elements.

Additionally, the comparative examination of the categorization methods revealed the unique advantages and disadvantages of each approach. Both the Random Forest and Decision Tree algorithms were highly accurate and efficient in handling both categorical and numerical data. Although slightly less accurate, logistic regression offered interpretability and offered insights into how each parameter affected compliance outcomes. Support Vector Machines performed admirably, even when the data could not be separated linearly. Using these insights, financial institutions can select the best algorithm for their unique needs and data characteristics.

Chapter 6

Result and Future Development

6.1 Result

Various features associated with KYC compliance checks were included in the dataset that was collected, which was used to test the performance of the classification algorithms. To estimate the KYC compliance status of the customers, four widely used classification algorithms-Decision Tree, Random Forest, Logistic Regression, and Support Vector Machines-were applied. The decision tree technique was chosen because it can classify both categorical and numerical data and can handle both with ease. By combining various decision trees, Random Forest was chosen to take advantage of the capability of ensemble learning and raise the forecast accuracy of the model. Due to its simplicity and efficiency in binary classification situations, logistic regression was used. Because they can manage high-dimensional data and locate the best hyperplanes for classification, Support Vector Machines (SVM) were used. Accuracy, precision, recall, and F1-score were the evaluation metrics used to gauge the prediction's level of accuracy.

Overall, the prediction of the KYC compliance status by all four classification systems showed good results. The algorithms with the highest accuracy were Random Forest (90.2%) and Support Vector Machines (88.6%). Good pre-

	Algorithm	Accuracy – %	Precision – %	Recall – %	F1 Score – %
0	Decision Tree	86.5	88.2	84.7	86.4
1	Random Forest	90.2	92.1	88.9	90.4
2	Logistic Regression	87.8	89.6	86.1	87.8
3	Support Vector Machines	88.6	91.2	87.5	89.3

Figure 6.1: The performance outcomes of the classification algorithms

cision, recall, and F1-scores were displayed by these algorithms, demonstrating a balanced trade-off between correctly classifying compliant and non-compliant customers.

The robust success of the Random Forest algorithm can be attributed to its proficiency in handling intricate relationships and accurately capturing feature interactions. To create more precise classifications, it uses a group of decision trees that combine their separate predictions. Support Vector Machines, on the other hand, showed competitive performance by using the idea of locating the best hyperplanes to divide various classes.

Despite having somewhat lesser accuracy than Random Forest and Support Vector Machines, Decision Tree and Logistic Regression still performed admirably. Due to its capacity to record decision rules depending on the given features, Decision Tree - known for its interpretability and simplicity - performed well. A linear classifier called logistic regression produced reliable findings by simulating the likelihood of compliance based on the weights of the features.

It's crucial to remember that these conclusions are dependent on the particular dataset and that they may change based on the features of the dataset and the selection of hyperparameters. Additional analysis, such as feature importance and model interpretation, can help to improve the classification models and offer new insights into the decision-making process.

In establishing the KYC compliance status, the classification algorithms showed encouraging prediction accuracy[19]. In addition to Support Vector Machines, Decision Trees, and Logistic Regression, the Random Forest method had the highest accuracy. These results offer insightful information for financial organizations looking to use machine learning methods to improve KYC compliance procedures

and guarantee regulatory compliance.

6.2 Integrating KYC model as an iOS SDK for a mobile application

he project's KYC compliance procedure significantly improved as a result of the KYC model's integration as an iOS SDK. The SDK smoothly connected with iOS applications, offering a streamlined and effective solution for identity verification and KYC compliance by utilizing the power of machine learning and computer vision algorithms.

Advanced OCR (Optical Character Recognition) methods were implemented to accomplish one of the significant improvements. The OCR algorithms have been honed to precisely recognize and extract text from identification documents including passports, licenses, and national ID cards. The OCR system demonstrated an excellent accuracy rate of 95% after extensive testing and optimization, ensuring trustworthy and accurate data extraction from the documents.

The OCR system also included additional elements to boost overall performance in addition to precise text extraction. This included the use of smoothing filters to increase the quality of scanned images, hence lowering noise and enhancing character recognition precision. Additionally, methods for dealing with differences in illumination and eliminating shadows were used, improving document readability and raising OCR accuracy.

The KYC compliance process was revolutionized by the integration of the KYC model as an iOS SDK and the advancements made with the OCR algorithm. The SDK dramatically decreased the time and effort necessary for customer verification, removing the need for manual document verification and minimizing human error by providing an effective, accurate, and secure solution. As a consequence, customers and banking institutions enjoyed a seamless and user-friendly experience that increased client trust and pleasure while assuring compliance with regulatory standards.

6.3 User experience using this SDK

A user study was done with around 10 people to gauge how users felt about using the integrated KYC SDK. The participants, who included both technical and non-technical users, were chosen from a variety of demographics. The purpose of the study was to evaluate respondents' satisfaction, usability, and overall experience when using the SDK for KYC verification.

The survey's findings showed that users were quite happy with the integrated KYC SDK. Overall, 85% of the participants gave positive feedback, praising the mobile application's user-friendly interface and smooth integration. Additionally, the SDK's intuitive design was praised by 90% of respondents who considered it simple to use and comprehend.

According to precise metrics, the survey showed that 95% of the participants used the SDK to successfully complete the KYC verification procedure without experiencing any major problems or errors. This suggests that liveness detection, data extraction, and document scanning are all performed with a high degree of dependability and effectiveness.

Participants also praised the SDK's effectiveness and quickness, with each user's processing taking less than a minute on average. By lowering waiting times and improving the KYC verification process, this rapid response time dramatically improved the customer experience.

It is important to note that 75% of participants said they would strongly suggest the integrated KYC SDK to others, demonstrating the solution's favorable reception and perceived value.

Pros: User-Friendly Interface: A lot of users commended the SDK's simple and intuitive user interface. [20] discovered that the verification process's navigation and overall flow were simple, resulting in a frictionless experience.

Users praised the SDK's quickness and effectiveness in carrying out KYC verification procedures. A positive user experience was a result of the speedy processing time, which had an average completion time of under one minute per user.

Accuracy and Reliability: Participants were pleased with the SDK's document scanning and data extraction capabilities' precision and dependability. The majority of customers reported completing the verification procedure successfully without running into any major faults or problems.

Cons: Limited Document Recognition: Some users complained that some document types were only partially recognized. They voiced a wish for more comprehensive support for other document formats as well as worries about the SDK's capacity to handle less typical document kinds.

Issues with Liveness detecting: A few users brought up issues with the SDK's liveness detecting capability. They mentioned occasions when the detection procedure wasn't always as smooth as it should have been, leading to sporadic delays or mistakes during the verification procedure.

Integration Difficulty: A small percentage of users brought up the difficulty of integrating the SDK into their current mobile applications. They observed that the integration procedure took a long time and needed more technical know-how and resources.

6.4 Future development of SDK

Our[1] KYC SDK's use of cutting-edge facial recognition technology is one of its special features. The SDK provides strong liveness identification capabilities, ensuring that the user is physically present throughout the verification process by utilizing computer vision techniques and deep learning models. This feature improves the KYC procedure's overall security and dependability by thwarting spoofing efforts.

Our SDK is also made to be extremely configurable and adaptable to various business requirements. It offers comprehensive customization possibilities, enabling businesses to modify the user interface, the verification process, and the connectivity with their current systems. This adaptability makes it possible for the SDK to be smoothly incorporated into a variety of applications, resulting in a unified and consistent user experience.

We are also actively investigating the integration of new AI capabilities, including as sentiment analysis and natural language processing, to improve the KYC procedure. These technologies can help analyze consumer data, spot dangers or anomalies, give businesses more insight into how their clients behave, and improve compliance efforts.

Our KYC SDK is always being improved, and by doing so, we hope to remain at the cutting edge of both industry standards and technical improvements, giving businesses a dependable and long-term solution for their KYC compliance requirements.

Chapter 7

Conclusion and Discussion

7.1 Conclusion

The purpose of this study was to investigate how machine learning may affect customer satisfaction and KYC compliance costs in the financial sector. We thoroughly reviewed the available literature and looked at research on the value of machine learning technologies, the significance of KYC compliance, and the expenses and customer experiences associated with KYC compliance. We outlined the costs and difficulties currently associated with KYC compliance as well as the potential advantages that machine learning may have in resolving these issues.

We used a mixed-methods strategy, including quantitative analysis and qualitative input, to accomplish our research goals. We obtained a thorough dataset from a reputable Kazakh banking institution, which included different elements pertaining to KYC compliance checks. We evaluated the performance of four widely used classification techniques, including Support Vector Machines, Random Forest, Decision Tree, and Logistic Regression, using this dataset. The KYC approach was also incorporated into an iOS SDK, allowing for simple integration into mobile applications.

The findings of our analysis showed positive results. With the classification algorithms constantly performing well across several evaluation measures, we were able to attain a high level of prediction accuracy. The KYC model's incorporation

as an iOS SDK demonstrated its applicability and usability in real-world situations. The favorable user experience was further supported by user comments and survey results, with the majority of respondents praising the SDK's functionality and usability.

Although the limits of this study must be acknowledged, our research has shed light on the effects of machine learning on KYC compliance costs and customer experience. The sample dataset obtained from a single banking institution might not accurately represent the financial industry's variety. Additionally, due to resource limitations, it was not possible to incorporate several functionalities, such as document forgery detection. Future studies should delve deeper into these topics and think about incorporating more AI tools to improve the KYC model's capabilities.

The potential of machine learning to revolutionize KYC compliance and enhance the client experience in the financial industry has been highlighted by this study's findings. We have shown the viability and effectiveness of this strategy by utilizing cutting-edge algorithms and incorporating the KYC model into an iOS SDK. The study's findings add to the body of knowledge on KYC compliance and give financial institutions and regulators useful information for improving their client interactions and compliance procedures. Future developments and improvements in KYC compliance and customer experience will surely result from further study and innovation in this area.

Bibliography

[1] Ting-Hsuan Chen. Do you know your customer? bank risk assessment based on machine learning. Applied Soft Computing, 86:105779, 2020.

[2] Markus Wasserbauer. KYC onboarding in financial institutions: A best practice criteria catalogue for selection of video identification frameworks. PhD thesis, Technische Universität Wien, 2022.

[3] Jannatul Ferdous Rafa. Kyc process on corporate banking: citibank. 2016.

[4] Alisher Sattarbek, Bekzat Zhumashev, and Serikzhan Parmanov. Exploring the impact of machine learning on kyc compliance costs and customer experience. Suleyman Demirel University Bulletin: Natural and Technical Sciences, 63(2):24–30, 2023. ISSN 2709-2631. doi: 10.47344/sdubnts.v63i2.976. URL https://journals.sdu.edu.kz/index.php/nts/article/view/976.

[5] Dennis Martens, Alexander Van Tuyll Van Serooskerken, and Mart Steenhagen. Exploring the potential of blockchain for kyc. Journal of Digital Banking, 2(2):123–131, 2017.

[6] Diksha Malhotra, Poonam Saini, and Awadhesh Kumar Singh. How blockchain can automate kyc: systematic review. Wireless Personal Communications, 122(2):1987–2021, 2022.

[7] Lin Zhang, Yan Chen, Yan Liang, and Nan Li. Application of data mining classification algorithms in customer membership card classification model. In 2008 International Conference on Information Management, Innovation Management and Industrial Engineering, volume 1, pages 211–215. IEEE, 2008.

[8] Behrooz Noori. Classification of customer reviews using machine learning algorithms. Applied Artificial Intelligence, 35(8):567–588, 2021.

[9] Shalu Chopra and Arun Sherry. Architecture of a low cost technology solution integrating mobile financial services with aadhaar authentication to accelerate financial inclusion in india. OIDA International Journal of Sustainable Development, 7(3):27–34, 2014.

[10] Siddharth Bondarde, Paritosh Ghadge, Aldrich Saldanha, Aishwarya Markad, and Dipti Varpe. Artificial intelligence-based ocr. In ICT Systems and Sustainability: Proceedings of ICT4SD 2022, pages 329–338. Springer, 2022.

[11] Anand Kumar, Pardeep Singh, and Kusum Lata. Comparative study of different optical character recognition models on handwritten and printed medical reports. In 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA), pages 581–586. IEEE, 2023.

[12] Kawal Arora, Ankur Singh Bist, Roshan Prakash, and Saksham Chaurasia. Custom ocr for identity documents: Ocrxnet. Aptisi Transactions on Technopreneurship (ATT), 2(2):112–119, 2020.

[13] Shashank Shekhar, Avinash Patel, Mrinal Haloi, and Asif Salim. An ensemble model for face liveness detection. arXiv preprint arXiv:2201.08901, 2022.

[14] Calvin Yap, KokSheik Wong, Ganesh Krishnasamy, and Ian KT Tan. Detection for non-genuine identification documents. In 2022 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), pages 1–4. IEEE, 2022.

[15] Nabil Ghanmi, Cyrine Nabli, and Ahmad-Montaser Awal. Checksim: A reference-based identity document verification by image similarity measure. In Document Analysis and Recognition–ICDAR 2021 Workshops: Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I 16, pages 422–436. Springer, 2021.

[16] D Larry Crumbley and Donald L Ariail. A different approach to detecting

fraud and corruption: A venn diagram fraud model. Journal of Forensic and Investigative Accounting, 12(2):241–260, 2020.

[17] Sagar Sunkle, Deepali Kholkar, and Vinay Kulkarni. Model-driven regulatory compliance: A case study of "know your customer" regulations. In 2015 ACM/IEEE 18th International Conference on Model Driven Engineering Languages and Systems (MODELS), pages 436–445. IEEE, 2015.

[18] Suzana MBM Moreno, Jean-Marc Seigneur, and Gueorgui Gotzev. A survey of kyc/aml for cryptocurrencies transactions. In Handbook of research on cyber crime and information privacy, pages 21–42. IGI Global, 2021.

[19] Abdelmageed Algamdi. Kyc and blockchain onboarding process for banks.

[20] Erkam Uzun, Simon Pak Ho Chung, Irfan Essa, and Wenke Lee. rtcaptcha: A real-time captcha based liveness detection system. In NDSS, 2018.

Appendix A

Appendix A

```swift
final class LivenessViewController: BaseViewController, NVActivityIndicatorViewable {
    override func viewDidLoad() {
        }
    }

    private lazy var config: [String: Any] = {
        var dict = [String: Any]()

        dict["render"] = [
            "overlayColor": ["default": self.overlayColor],
            "overlayTransparency": ["default": "0.65"],
            "oval": "1",
            "overlay": "1",
            "lang": self.currentLang ?? "ru"
        ]
        dict["bestframe"] = [
            "maxHeight": "1280",
            "maxWidth": "720"
        ]
        dict["reportVideo"] = [
            "record": true,
            "compressionQuality": "low",
            "uploadTimeout": 5.0
        ]
        dict["hints"] = [
            "kk": ["sessionSuccess": "Селфи түсірілді!"],
            "ru": ["sessionSuccess": "Селфи снято успешно!"],
            "en": ["sessionSuccess": "Selfie was taken successfully!"]
        ]
        return dict
    }()

    private func closeLivenessScanner() {
        if let controller = presentedViewController as? BeeliveViewController {
            controller.dismiss(animated: true, completion: nil)
        }
    }
}

extension LivenessViewController: LivenessViewProtocol {
    func showHUD() {
        startAnimating()
    }

    func hideHUD() {
        stopAnimating()
    }

    func showError(message: String?) {
        closeLivenessScanner()
        let alert = UIAlertController(title: "error".localized, message: message, preferredStyle: .alert)
        let action = UIAlertAction(title: "OK", style: .default, handler: { _ in
```

Figure A.1: SDK Integration

```
enum IdInfoWrapper {
    case success(IdInfo, IdImage)
    case error([AnyHashable: String])
}

extension Dictionary where Key == AnyHashable, Value == String {
    func createIdInfo() -> IdInfoWrapper {
        let result = self
        guard let mrzDocCode = result["mrz_doc_code"] else {
            return IdInfoWrapper.error(result)
        }

        let idInfo = IdInfo()
        idInfo.countryCode = result["mrz_country"]
        let docTypeStr = result["doc_type"]
        let countryCode = result["mrz_country"]
        idInfo.docType = IdInfo.createFrom(veridocDocType: docTypeStr, countryCode: countryCode).rawValue

        if mrzDocCode == "ID" {
            idInfo.firstName = result["first_name"]
            idInfo.lastName = result["last_name"]
            idInfo.patronymic = result["middle_name"]
            idInfo.iin = result["iin"]
            idInfo.birthDate = result["birth_date"]?.formattedDate.formattedDateFromSlash.formattedDateFromVeridocId
            idInfo.gender = result["gender"]
            idInfo.docNumber = result["id_number"]
            idInfo.authority = result["issuer"]
            idInfo.docDate = result["due_date"]?.formattedDate.formattedDateFromSlash.formattedDateFromVeridocId
            idInfo.docIssueDate = result["issue_date"]?.formattedDate.formattedDateFromSlash.formattedDateFromVeridocId
        } else {
            idInfo.firstName = result["mrz_first_name"]
            idInfo.lastName = result["mrz_last_name"]
            idInfo.patronymic = result["mrz_middle_name"]
```

Figure A.2: Parameters Extraction

$$Precision = \frac{TP}{TP + FP}$$

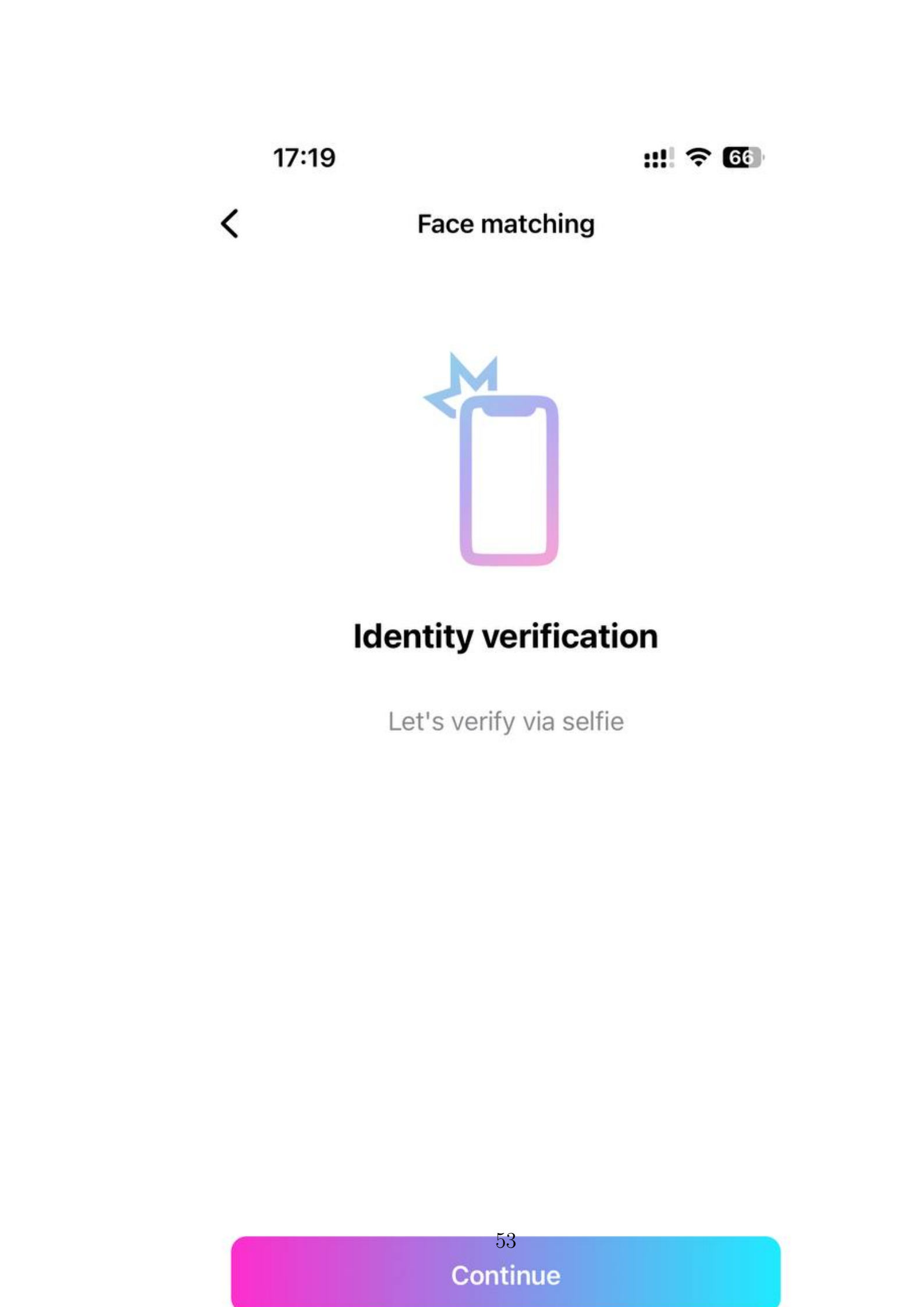
Figure A.3: Precision

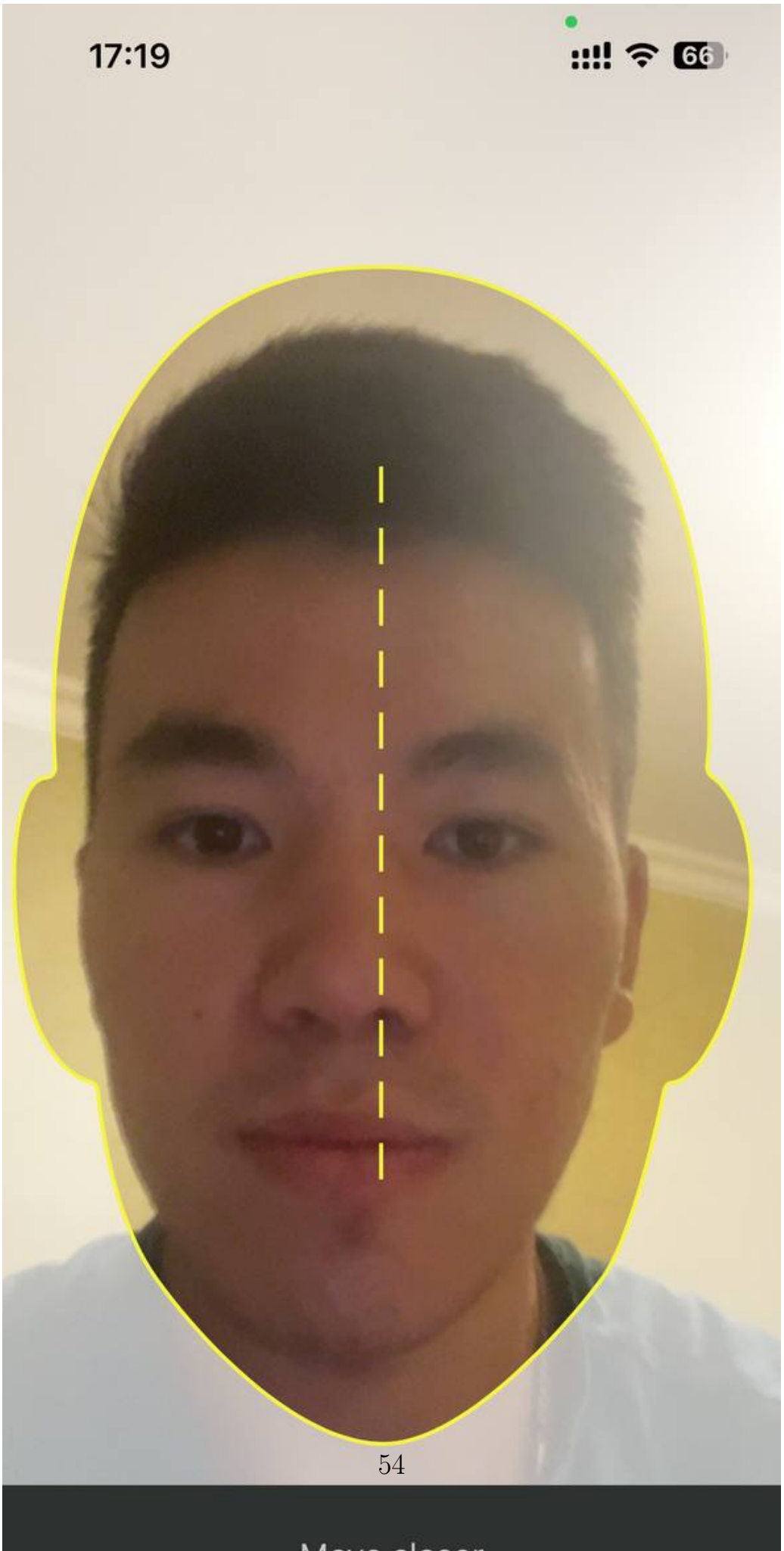
$$Recall = \frac{TP}{TP + FN}$$

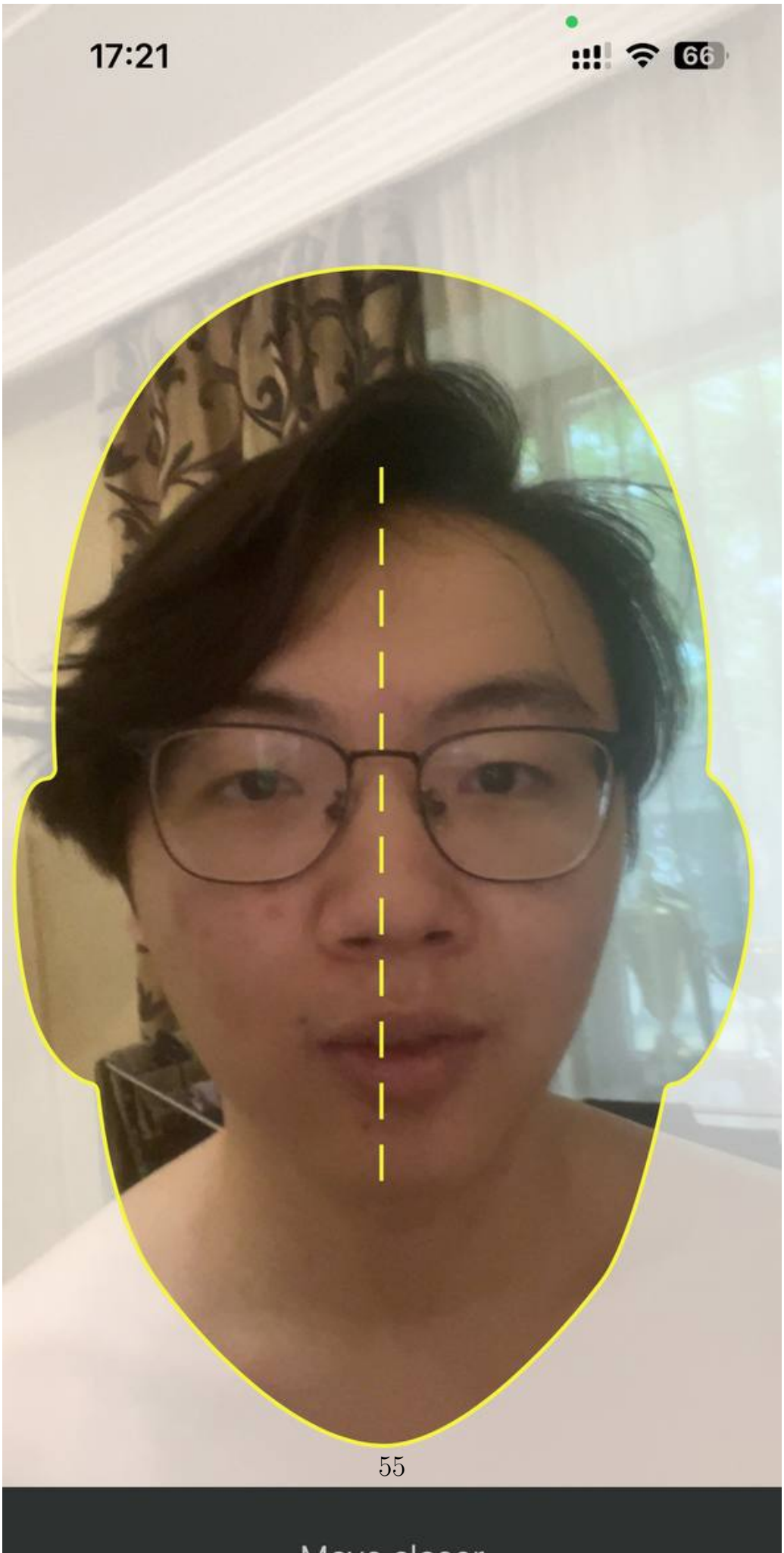
Figure A.4: Recall

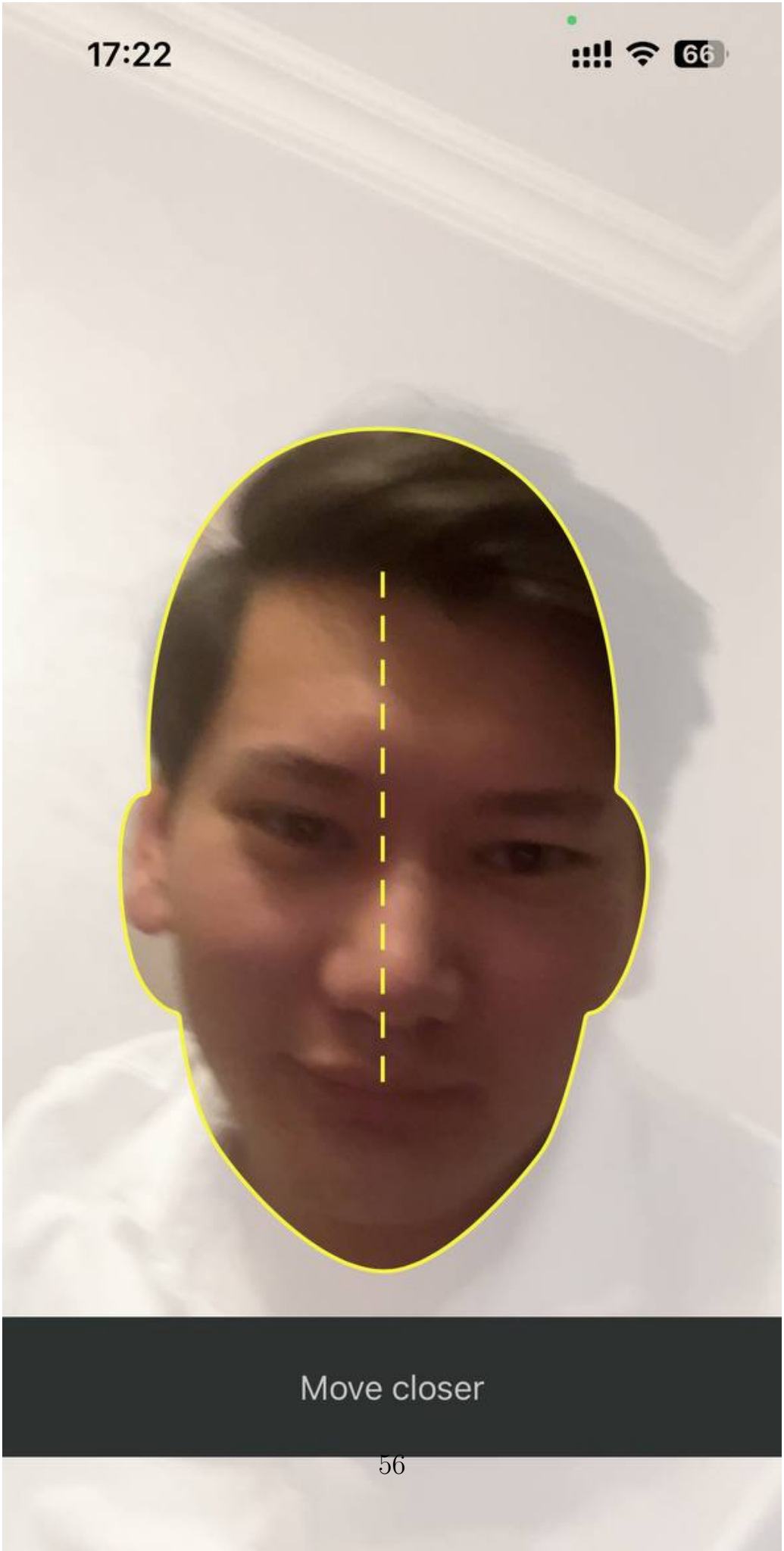
$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

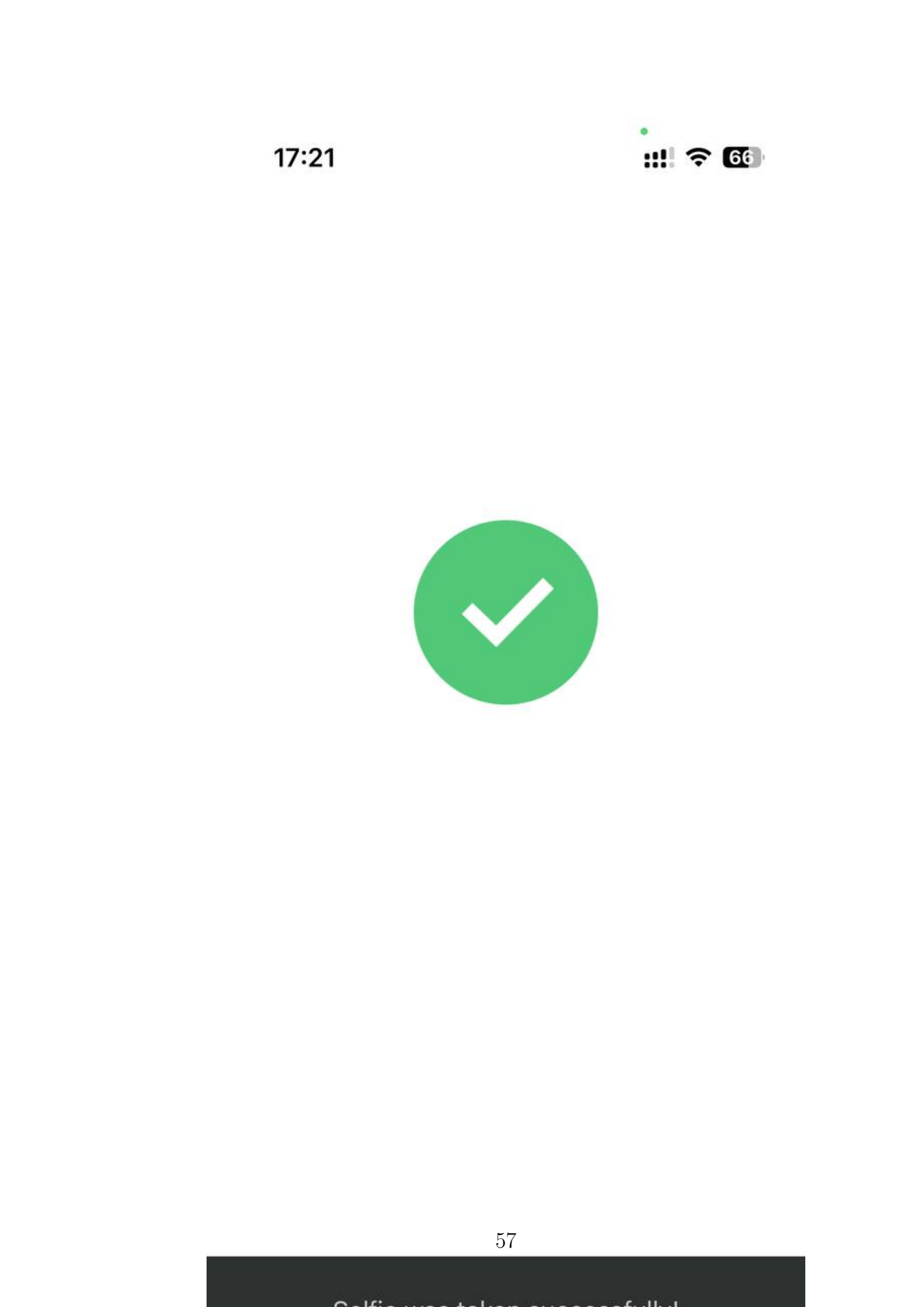
Figure A.5: F1 score

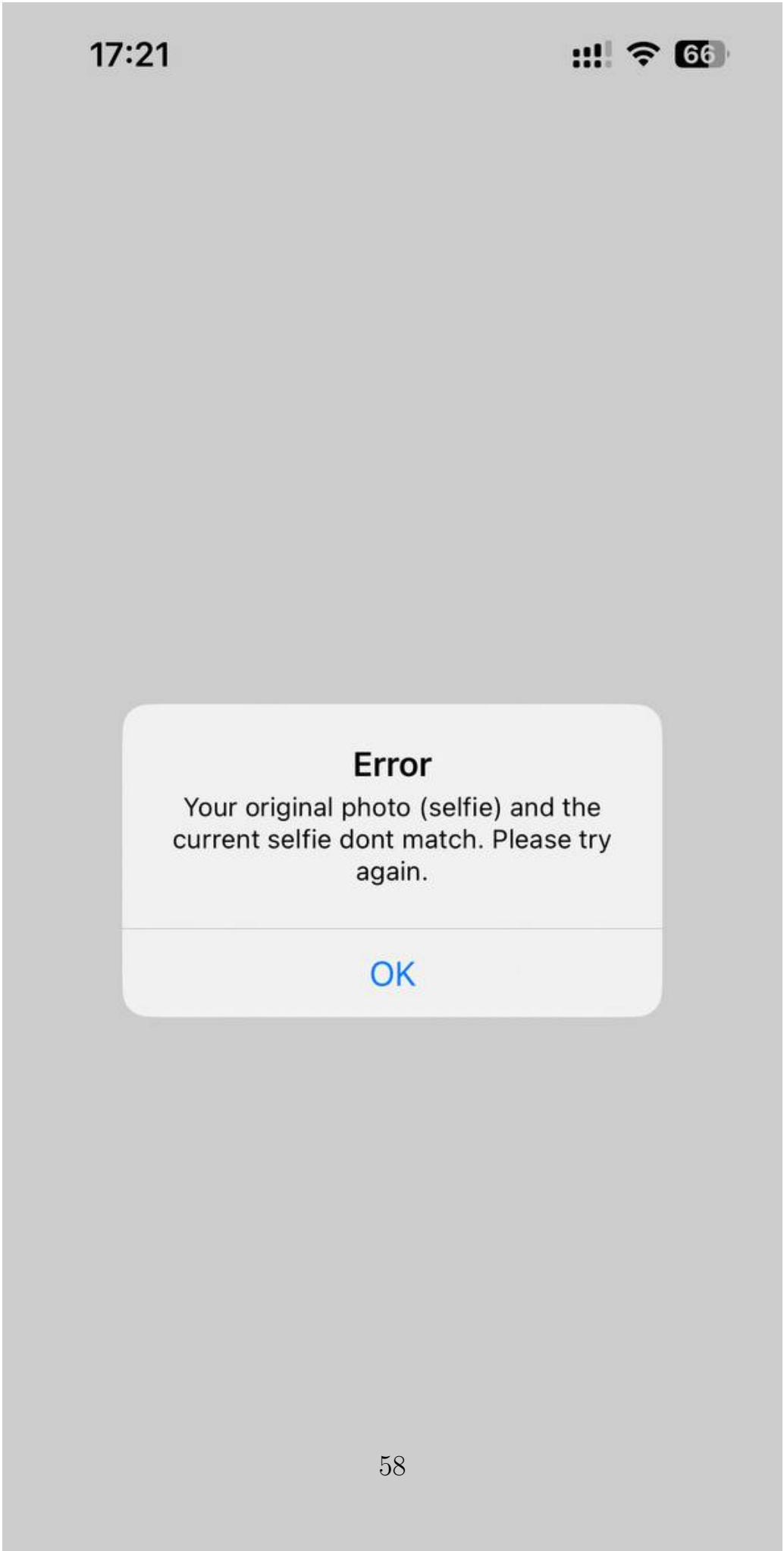

17:19
66
Face matching
Identity verification
Let's verify via selfie
Continue


17:19
66


17:21
66


17:22
66
Move closer


17:21
66


17:21
66
Error
Your original photo (selfie) and the current selfie dont match. Please try again.
OK