Streamlining Tax Calculations: An Automated Approach for Individual Taxpayer

Sahil Saurabh Jain[1], Anuj Kumar Bharti[1], Swati Sharma[2]

Bennett University, Gautam Buddh Nagar India [1], Galgotias University, India[2]

Corresponding author: Anuj Kumar Bharti, Email: bharti.anujk@gmail.com

In this paper, we provide an innovative tax computation method that intends to automate and simplify the calculation of taxes for Indian citizens. The method that we suggest in this work uses machine learning (ML), natural language processing (NLP), optical character recognition (OCR), and OCR technologies to accurately interpret financial data from a user's bank statement. The method we provide here leverages and employs data extraction, cleaning, sorting, and categorization using NLP techniques, as well as tax computation based on the current Indian tax framework. Our approach tackles both the shortcomings of current automated tax computation systems as well as the difficulties Indian taxpayers encounter when navigating the complicated tax structure of their nation. With the help of automation of the tax calculation process, we aim to reduce the likelihood of errors and minimize the time and effort required by taxpayers while preventing fraudulent activities. Although our approach has shown promising results, there are limitations and areas for further development, such as improving OCR accuracy, adding support for different languages, managing complex tax scenarios (for people owning business), enhancing scalability, and addressing privacy and security concerns. With further research and development, our proposed tax computation method has the potential to streamline and enhance the tax filing process in India, making it more efficient, accurate, and user-friendly.

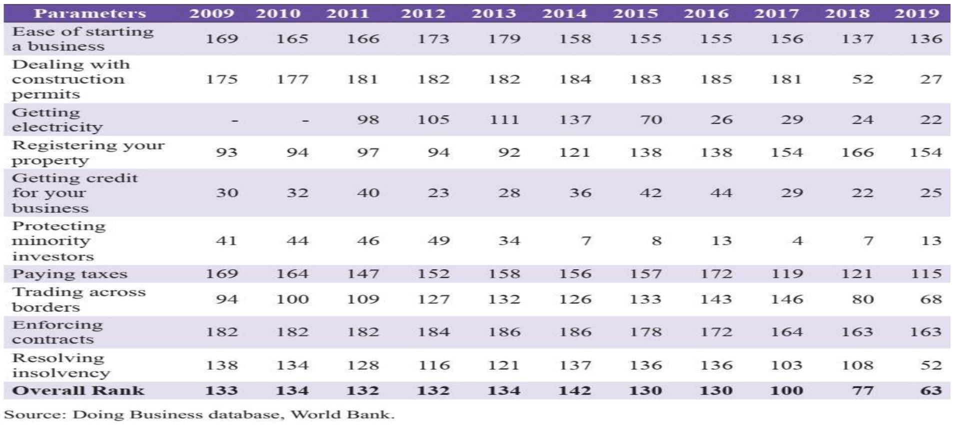
Keywords: OCR, Tax calculation, Natural Language Processing, Machine Learning.

2023. In Satyasai Jagannath Nanda & Rajendra Prasad Yadav (eds.), *Data Science and Intelligent Computing Techniques*, 401–410. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.56155/978-81-955020-2-8-36

1 Introduction

India has emerged as the world's fastest expanding economy, and with this rapid expansion has come a painfully complicated tax system. Filing taxes is an integral part of financial planning for any individual or business. In India, navigating through the process of filing taxes can be very complex, with various forms to be filled, documents to be submitted and rules to be followed. According to a yearly ‘Ease of Doing Business’ report by World Bank, India ranks 115th out of 190 countries [1] in the ease of paying taxes, highlighting the country’s painfully complex tax system. Figure 1 depicts ranking of India on various parameters of doing business according to world bank.The Indian government is very aware of this problem as it has been taking steps to simplify and ease the tax filing process. Income Tax Department of India, keeping up with digitization of the nation, launched the electronic tax filing system of Income Tax Returns because the lion’s share of revenue of the country is generated by direct taxes. E-taxation scheme was one of the “action lines” introduced in Indian tax machinery in the A.Y. 2006-07 for all assessments for improving the Return filing system [2]. The past year Indian government also increased the threshold for tax exemption, an attempt to reduce the burden on lower income households. Despite these efforts by the government to simplify the process, filing taxes remains a daunting task for many individuals and businesses alike. The process of filing taxes involves gathering and organizing various documents such as bank statements, rent receipts, and salary receipts. The task of categorizing and extracting relevant information from these documents can be time-consuming and prone to errors. Additionally, fraudulent activities such as misrepresentation of income or false deductions can lead to incorrect tax calculations. These challenges make tax filing a complicated task for individuals and businesses, leading to confusion and delays in the process. Automating the process of tax calculation using technology can help alleviate the burden and make it easier for people to file their taxes.

A study conducted for the IMF in 2014 provides some insightful analysis of the challenges faced by modern tax administrations in achieving optimal tax collections while minimizing administration and compliance costs [3]. The authors of this working paper argued that self-assessment is the most effective means of achieving voluntary compliance, and that tax administrations need to consistently apply self-assessment principles in their income tax laws. This recommendation is consistent with a 2019 study conducted by Atlanta press, which also emphasized the importance of self-assessment in the digital age of tax administration [4]. However, the authors note that many tax administrations continue to rely heavily on desk auditing, with risk management practices being largely underdeveloped or underutilized.



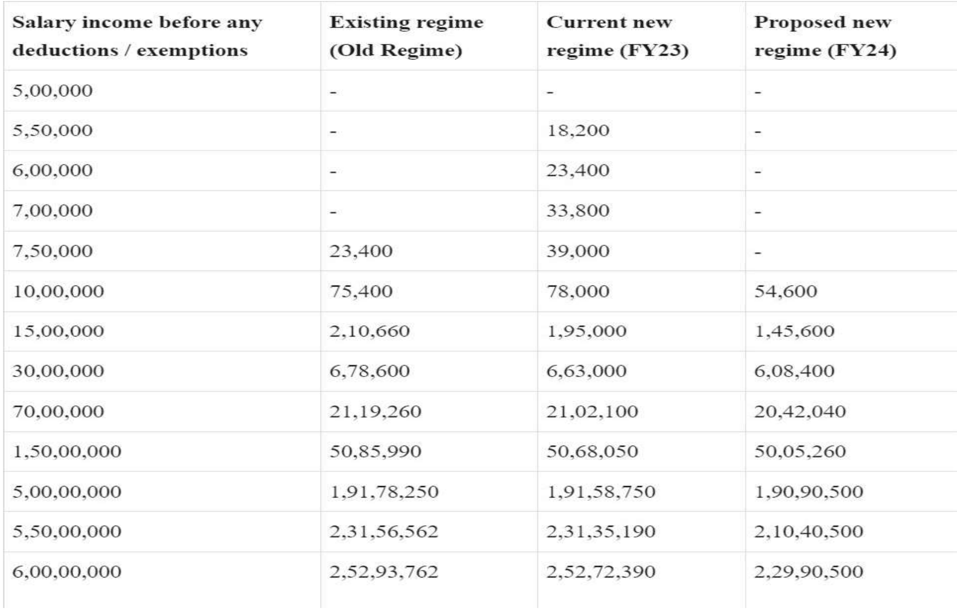
Parameters	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Ease of starting a business	169	165	166	173	179	158	155	155	156	137	136
Dealing with construction permits	175	177	181	182	182	184	183	185	181	52	27
Getting electricity	-	-	98	105	111	137	70	26	29	24	22
Registering your property	93	94	97	94	92	121	138	138	154	166	154
Getting credit for your business	30	32	40	23	28	36	42	44	29	22	25
Protecting minority investors	41	44	46	49	34	7	8	13	4	7	13
Paying taxes	169	164	147	152	158	156	157	172	119	121	115
Trading across borders	94	100	109	127	132	126	133	143	146	80	68
Enforcing contracts	182	182	182	184	186	186	178	172	164	163	163
Resolving insolvency	138	134	128	116	121	137	136	136	103	108	52
Overall Rank	**133**	**134**	**132**	**132**	**134**	**142**	**130**	**130**	**100**	**77**	**63**

Source: Doing Business database, World Bank.

Figure 1: Capturing India's decade long journey in the Doing Business rankings

Despite the growing popularity of automated tax computation systems, there are currently few real-world solutions that can accurately and swiftly compute taxes using relevant financial facts. Several modern solutions created by companies like Groww, Clear, and TurboTax rely on manually input data, which can be time-consuming and error prone. Automated tax systems or online tax systems have also been introduced in a number of nations; similar to previous examples they rely on manually input data and place a greater emphasis on data discrepancies and tax fraud detection than on simplifying taxpayers' tax filing processes [5]. The study article Automated Tax Return Verification using Blockchain Technology suggests using Blockchain technology as a way to automate tax verification in Bangladesh, a nation with a tax filing system comparable to India's and one that has similar problems [6]. On the other end a paper Tax compliance and privacy rights in profiling and automated decision making argued about how automated decision making in tax matters are included in the broader public interest exception, safeguards to taxpayers' privacy rights need to be in place [7].

Google has also made an effort to address this issue by creating Document AI. Compared to conventional approaches, document AI has a number of benefits, including quicker processing times, more accuracy, and a lower chance of fraud. Invoices, receipts, and contracts are just a few examples of the sorts of data that may be extracted and categorized using document AI. This data can then be utilized for a variety of tasks, including tax compliance, auditing, and financial analysis. Nevertheless, there are also drawbacks to using Document AI, including the requirement for a sizable quantity of training data, the possibility of bias in the training data, and the possibility of OCR mistakes [8].



Salary income before any deductions / exemptions	Existing regime (Old Regime)	Current new regime (FY23)	Proposed new regime (FY24)
5,00,000	-	-	-
5,50,000	-	18,200	-
6,00,000	-	23,400	-
7,00,000	-	33,800	-
7,50,000	23,400	39,000	-
10,00,000	75,400	78,000	54,600
15,00,000	2,10,660	1,95,000	1,45,600
30,00,000	6,78,600	6,63,000	6,08,400
70,00,000	21,19,260	21,02,100	20,42,040
1,50,00,000	50,85,990	50,68,050	50,05,260
5,00,00,000	1,91,78,250	1,91,58,750	1,90,90,500
5,50,00,000	2,31,56,562	2,31,35,190	2,10,40,500
6,00,00,000	2,52,93,762	2,52,72,390	2,29,90,500

Figure 2: India's New tax Regime

In order to overcome these difficulties, we combine machine learning (ML) and natural language processing (NLP) techniques to create a novel strategy for autonomously calculating taxes in India. We will start by using a method known as document scan categorization because we will be working with paper-based data. Document scan classification is a way of categorizing scanned documents based on their content. Large numbers of scanned documents may be organized, saved, and identified using the classification process. We will utilize OCR, or optical character recognition, to implement this approach. OCR may be used to scan text documents and aid in text categorization by extracting text from the documents and applying it for text classification. With OCR technology, handwritten or printed text may be transformed into machine-encoded text. OCR has been used extensively for many years in a variety of businesses, primarily for digitizing text documents. In one study, Javier Ferrando (2020) employed OCR to extract text from scanned documents, and the obtained text was subsequently used for document classification. The study showed an accuracy rate of 89.47 when classifying the text using EfficientNet models [9].

The required financial information may, however, be difficult to extract from these papers due to the possibility of unstructured writing. Using NLP approaches, this problem may be solved by automatically identifying the texts and extracting the pertinent financial data. Natural Language Processing (NLP) is an area of artificial intelligence that deals with the interaction of computers and humans through natural language. By utilizing several strategies and techniques including Named Entity Recognition, Sentiment Analysis, and Information Extraction, NLP may be utilized to categorize documents and extract pertinent financial data that can aid in computing an individual's tax [10].

In India, particularly for individuals and small businesses with limited financial resources, we believe that our recommended method has the potential to significantly improve tax computation efficiency and accuracy. Since Indian tax system is designed for covering whole population by having multiple tax slabs and multiple caveats. Figure 2 shows the latest tax regime of Indian government. Hence, by automating the tax computing process and implementing fraud detection tools, we can decrease the likelihood of errors and fraud while simultaneously lowering the time and effort required by taxpayers. Furthermore, by combining OCR, supervised learning, natural language processing, and machine learning technologies, tax computations may be done more precisely, and personalized advise based on each person's unique financial situation may be offered.

2 Methods

2.1 Data Extraction

The first step in the proposed methodology is data extraction. The methodology begins by taking the user's bank statements in PDF format. PDF files are the most commonly used format for bank statements. Tablo is a tool that helps to extract tables from each page of the PDF. In this research, we use Tablo to extract the tables from each of the pages in the PDF [11].

Once we have extracted the tables from each page, we convert each table into a Pandas data frame. Pandas is an open-source data manipulation and analysis library that is widely used in data science. The data frames are then merged into one consolidated data frame. We adjust all the serial numbers so that they are in order to avoid any inconsistencies in the final output. This consolidated data frame is the input for the next step. Figure 3 shows the automated tax computation using proposed system.

2.2 Data Cleaning and Sorting

Data cleansing and sorting comes next. The condensed data frame is sorted to only show the transactions from the selected year. We take transactions from April 1 of the selected year through March 31 of the following year. This process aids in data simplification and facilitates the computation of income tax.

The next stage is to merge the deposit and withdrawal parts into a single section and give deposits a positive value and withdrawals a negative one. This process aids in data simplification and facilitates the computation of income tax. Then, using data cleaning techniques, we eliminate any inconsistent, duplicate, or missing values from the data. This guarantees the consistency and accuracy of the data.

2.3 NLP for Categorization

The next crucial stage in automating a person's income tax is the categorization of the data. In this stage, the values in the data-frame are categorized using natural language processing (NLP) methods. The study of the relationship between computers and human language is known as natural language processing (NLP). It is an essential tool for analyzing unstructured text data since it enables machines

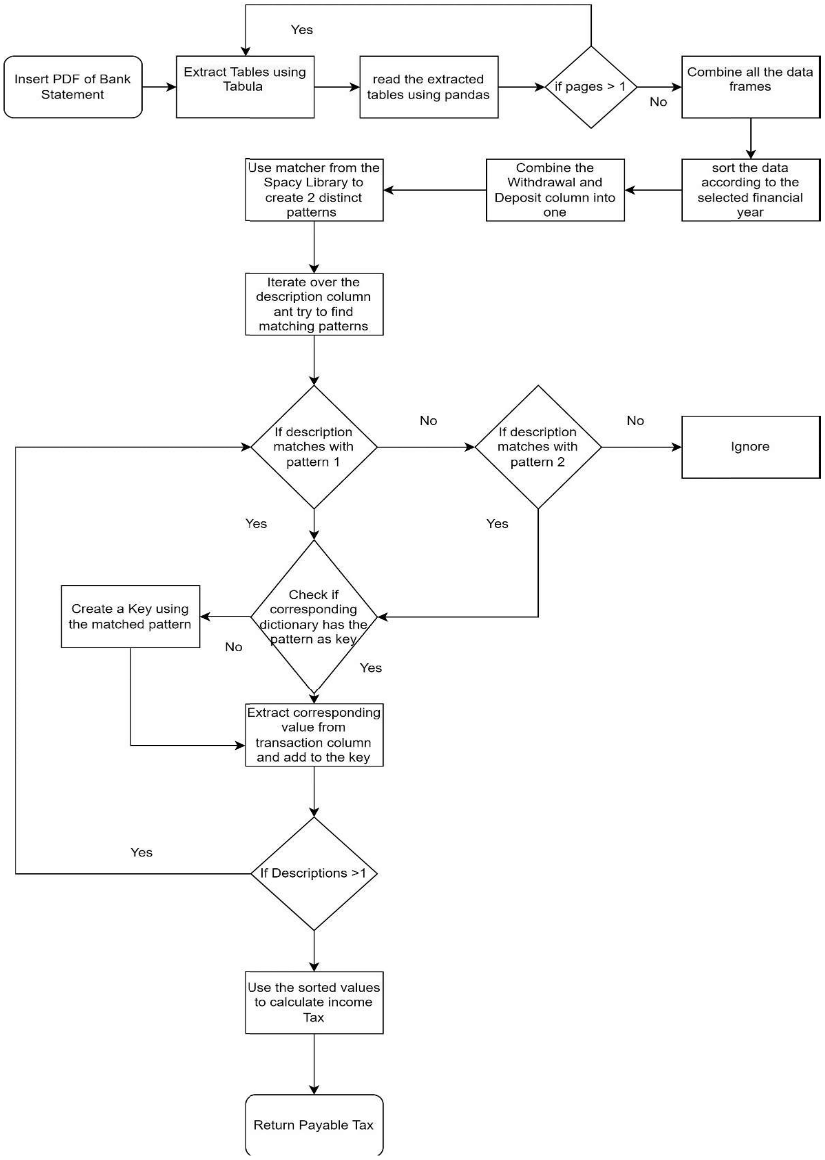

Yes
Insert PDF of Bank Statement
Extract Tables using Tabula
read the extracted tables using pandas
If pages > 1
No
Combine all the data frames
sort the data according to the selected financial year
Combine the Withdrawal and Deposit column into one
Use matcher from the Spacy Library to create 2 distinct patterns
Iterate over the description column ant try to find matching patterns
No
No
If description matches with pattern 1
If description matches with pattern 2
Ignore
Yes
Yes
Create a Key using the matched pattern
Check if corresponding dictionary has the pattern as key
No
Yes
Extract corresponding value from transaction column and add to the key
Yes
If Descriptions >1
Use the sorted values to calculate income Tax
Return Payable Tax

Figure 3: Flow Chart of Methodology

to understand and interpret human language. NLP techniques may be used for a range of applications, including sentiment analysis, chat-bots, and machine translation. They are used to extract meaning from natural language text input.
In this study, we use NLP to analyze the description part of the data frame. The description section contains text data that provides additional information about each transaction. This information can be used to identify the values that should be taken into account when calculating an individual's income tax.

We establish a dictionary of keywords and their categories in order tocategorize the data. We make a category for all the income sources an individual may have and all the expenditure that is exempted from the income tax. The necessary values in the data frame's description section are then found using the dictionary. To match the keywords to the appropriate values, we utilize matcher, a utility offered by the spacy library. This method aids in classifying the values into income and expenditure and separating taxable from non-taxable income.

After categorizing the values, we retrieve the matching values from the data frame's 'Overall' column. This phase assists us in distinguishing between the values that should be considered when computing an individual's income tax and those that should be omitted. With this procedure, we are able to separate the crucial data from the unimportant transactions.

2.4 Tax Calculation

We utilize a standard Python script to compute the tax payable based on the classification of the data in the last phase of automating an individual's income tax. This phase is adding all of the positive numbers for income and subtracting all of the negative values for expenses that we have classified for income tax purposes.

Once the data has been classified, we must apply the relevant tax rates based on government laws to compute the tax owed. The method used to calculate taxes varies by country. In India, for example, there are multiple tax slabs based on an individual's annual income, and each slab has a distinct tax rate. The government sets the tax rates, which are subject to vary from year to year. With slight tweaks to the tax computation method, the recommended approach of automating income tax calculation may be utilized in any nation. But, for the sake of this research, we will concentrate on India's tax system.

In India, the tax computation formula takes into account the income earned during a financial year, which runs from April 1 to March 31 of the next year. The income is divided into different tax slabs, and each slab has a different tax rate. The tax rates range from 0\% to 30\% and are determined based on the income earned by an individual [12].

Once we have calculated the taxable income, we apply the relevant tax rate to calculate the tax due. The tax due is then subtracted from any TDS (Tax Deducted at Source) paid during the year to arrive at the final tax payable. Figure 4 shows the user-friendly tax calculation system using proposed method.

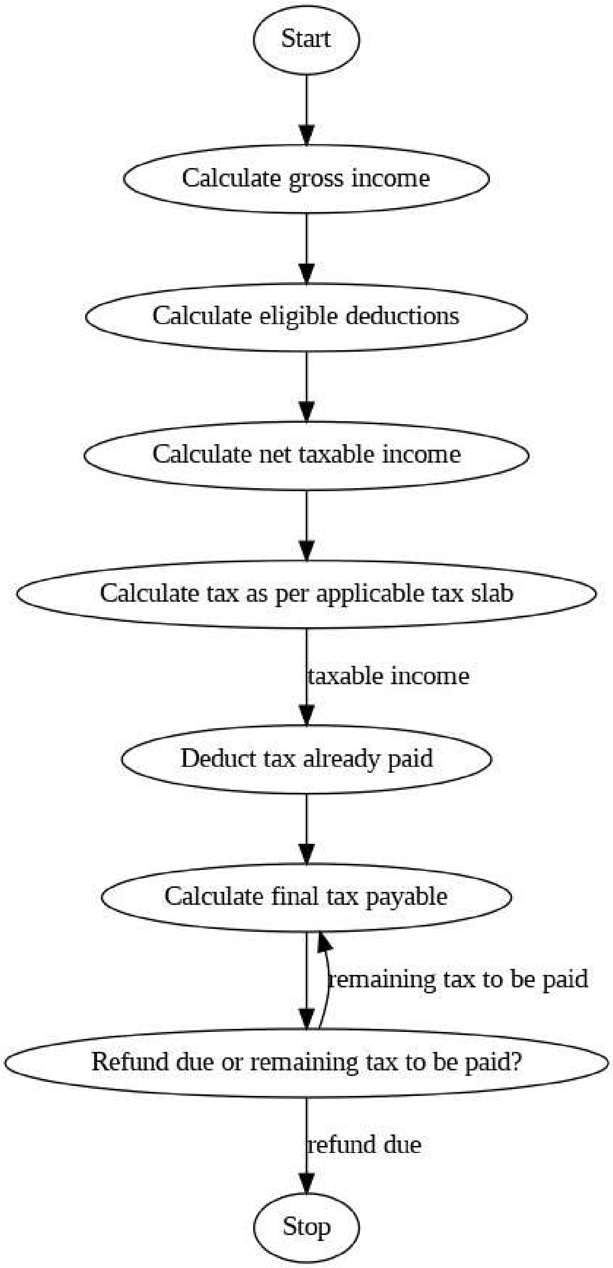

Start
Calculate gross income
Calculate eligible deductions
Calculate net taxable income
Calculate tax as per applicable tax slab
taxable income
Deduct tax already paid
Calculate final tax payable
remaining tax to be paid
Refund due or remaining tax to be paid?
refund due
Stop

Figure 4: Flow Chart of Tax Calculation

3 Results

In this study, we aimed to develop a system that simplifies and automates the process of calculating taxes in India using machine learning, natural language processing, and optical character recognition techniques. Our methodology involved data extraction, data cleaning and sorting, NLP for categorization, and tax calculation based on the Indian tax system.

The proposed system was successful in extracting relevant financial data from bank statements in PDF format using Tablo and converting the extracted data into a consolidated Pandas data frame. Data cleaning and sorting techniques were applied to filter the data and focus on transactions from the selected financial year, simplifying the tax computation process.

We employed a matcher, an NLP technique, in conjunction with pre-defined word pool(200 words) to analyze the description section of the data frame. Furthermore, we emphasized on using String Matching and Regular Expression Matching provided by the spacy library. This involved categorizing transactions as income or expenditure and distinguishing taxable from non-taxable transactions. Our approach was facilitated by a dictionary of keywords and their associated categories, ensuring precise categorization of financial transactions.

Finally, taking into account various tax slabs and rates in accordance with Indian governmental laws, we determined the tax due based on the categorized data and the current tax system. By making slight modifications to the tax computing process, this strategy can be utilized for other nations.

Our research demonstrates that the suggested approach can in theory greatly increase the accuracy and efficiency of tax computation in India, especially for low-income individuals. We can lessen the chance of errors and fraud while also saving taxpayers time and effort by automating the tax computing process and introducing fraud detection systems.

Moreover, the combination of OCR, supervised learning, natural language processing, and machine learning technologies enables more precise tax computations and personalized advice based on each individual's unique financial situation. This system can be further improved and scaled by incorporating additional data sources and addressing potential biases and OCR errors in the training data.

In conclusion, our study presents a possible approach for simplifying and automating the calculation of taxes in India, addressing the complexity as well as challenges that citizens in the nation experience. The broad use of this system and substantial advancements in the tax filing procedure may result from more study and development, which will help create a more effective and user-friendly tax system.

4 Limitations

Despite the promising results and potential benefits of the proposed tax computation system, there are several limitations that should be acknowledged:

- Data quality: The accuracy of the system depends on the quality of the input data. Inaccurate, incomplete, or inconsistent bank statements may negatively impact the performance of the system, leading to incorrect tax calculations.
- OCR errors: Optical character recognition (OCR) technology can sometimes misinterpret or fail to recognize characters in scanned documents, resulting in errors during the data extraction process. These errors may propagate through the subsequent stages of the system, affecting the accuracy of the tax computations.

- Language limitations: The current system is designed to process bank statements in English, which may not be applicable to regions with a diverse range of languages. Adapting the system to handle multiple languages would require additional efforts and resources.
- Privacy and security concerns: The system relies on accessing and processing sensitive financial data, which raises privacy and security concerns. Ensuring the protection of user data and maintaining compliance with data protection regulations are crucial challenges that need to be addressed in the development of the system [13].
- Changes in tax regulations: Tax regulations are subject to change over time, and the system must be regularly updated to reflect these changes to ensure accurate tax computations. This requires ongoing maintenance and updating of the system, which can be time-consuming and resource-intensive.
- Adoption barriers: Users might be hesitant to adopt the proposed system due to concerns regarding its accuracy, security, or complexity. Addressing these concerns and promoting user trust will be essential for widespread adoption of the system.

While the system does offer an advantageous way to improve the efficiency and precision of tax computations in India, it can be stated after weighing the advantages and disadvantages of the recommended tax calculation technique. For the system to continue to evolve and progress, it is essential to admit that a few issues need to be resolved. For example, the system might not be able to handle extremely complicated tax concerns or exceptions, which could result in mistakes or inaccurate information. Additionally, it may be difficult for some stakeholders to contribute the required funds for the system's infrastructure, software development, and training. Therefore, while the suggested tax computation system shows promise, it is essential to carefully weigh its benefits and drawbacks, and to continuously monitor and improve the system to meet the evolving needs of the tax system in India.

5 Future Work

The proposed tax computation system has demonstrated promising results in simplifying and automating the tax calculation process in India. To further improve the system and address its limitations, future work can focus on the following areas:

- Enhancing OCR accuracy: Improve the OCR technology used for data extraction by incorporating more advanced algorithms, increasing training data diversity, and incorporating error correction mechanisms to minimize errors during the data extraction phase. Shifting from Tabula to Camelot another python library that proves better at recognizing tables as it using advances computer vision.

- Multilingual support: Extend the system to support multiple languages by incorporating multilingual NLP models and adapting the data extraction process to handle bank statements in various languages, catering to a more diverse range of taxpayers.

- Scalability improvements: Optimize the system's performance to handle large volumes of data and process a high number of tax computations simultaneously, ensuring fast and efficient tax calculations for a broader user base.

- Privacy and security enhancements: Implement robust privacy and security measures to protect sensitive financial data, including data encryption, access controls, and secure storage solutions. Ensure compliance with data protection regulations and maintain user trust in the system.

- Adaptive tax regulation updates: Develop a mechanism to automatically update the system with the latest tax regulations, minimizing the need for manual intervention and ensuring that the system remains current with changing tax laws.

- User interface and user experience improvements: Design an intuitive and user-friendly interface for the system, making it more accessible to individuals and businesses with varying levels of tax knowledge and expertise.

Integration with existing tax software and platforms: Explore opportunities to integrate the proposed system with existing tax software or government platforms, allowing for seamless data exchange and improved interoperability. Real-world testing and validation: Conduct extensive real-world testing of the system, including trials with individuals and businesses, to validate its performance and identify areas for improvement. Collect feedback from users to refine the system and better understand their needs and expectations. By addressing these areas in future work, the proposed tax computation system can be further enhanced and refined, ultimately contributing to a more efficient, accurate, and user-friendly tax filing process for individuals and businesses in India and beyond.

References

[1] World Bank (2023). Ease of doing business. https://data.worldbank.org/indicator/IC.BUS. EASE.XQ,

[2] Arora, J, (2016). E-filing of income tax returns in india – an overviewe-filing of income tax returns in india – an overview. Scholarly Research Journal For Humanity Science English Language, 3(14):3434–3442

[3] Okello, A, (2014) Managing income tax compliance through self-assessment. International Monetary Fund.

[4] Jiang,B. (2020). Research on the application of big data technology in tax collection and management. Advances in Economics, Business and Management Research, 117:443–447,

[5] Aidonojie,P.A.,Nwazi, J. andEruteya, U. (2022). The legality, prospect, and challenges of adopting automated personal income tax by states in nigeria: A facile study of edo state. Cogito, 14(2):64–87,

[6] Hossain, S. Saha, S.,Akhi, J.F. andHelaly, T. (2020). Automated tax return verification with blockchain technology. In Proceedings of International Joint Conference on Computational Intelligence: IJCCI 2019, pages 45–55.

[7] Scarcella, L. (2019). Tax compliance and privacy rights in profiling and automated decision making. Internet Policy Review, 8(4),

[8] Hegghammer, T. (2022). Ocr with tesseract, amazon textract, and google document ai: a benchmarking experiment. Journal of Computational Social Science, 5(1):861–882

[9] Ferrando, J.,Domínguez, J.L.,Torres, J.,García, R., García, D.,Garrido, D., Cortada, J., and Valero, M. (2020). Improving accuracy and speeding up document image classification through parallel systems. In Computational Science–ICCS 2020: 20th International Conference, Amsterdam, The Netherlands, June 3–5, 2020, Proceedings, Part II 20, pages 387–400. Springer,

[10] Milana, C., and Ashta, A (2021). Artificial intelligence techniques in finance and financial markets: a survey of the literature. Strategic Change, 30(3):189–209,

[11] Mittal, R., and Garg, A., (2020). Text extraction using ocr: a systematic review. In 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), pages 357–362. IEEE,

[12] Income Tax Department. Income tax department, government of india. https://www. incometax.gov.in/, 2023.

[13] Zhou, L., (2009). Opportunities and challenges of artificial intelligence in the application of taxation system. In Proceedings of the 2019 International Conference on Economic Management and Cultural Industry (ICEMCI 2019), pages 201–206. Atlantis Press.