Deep Learning-based Delinquent Taxpayer Prediction: A Scientific Administrative Approach

YongHyun Lee[1] and Eunchan Kim[2,3*]
1 Department of Computer Science and Engineering, Seoul National University
Seoul 08826, Republic of Korea
[e-mail: leeyh@idb.snu.ac.kr]
2 College of Business Administration, Seoul National University
Seoul 08826, Republic of Korea
[e-mail: eunchan@snu.ac.kr]
3 Department of Intelligence and Information, Seoul National University
Seoul 08826, Republic of Korea
*Corresponding author: Eunchan Kim

Received September 3, 2023; revised November 30, 2023; accepted December 25, 2023; published January 31, 2024

Abstract

This study introduces an effective method for predicting individual local tax delinquencies using prevalent machine learning and deep learning algorithms. The evaluation of credit risk holds great significance in the financial realm, impacting both companies and individuals. While credit risk prediction has been explored using statistical and machine learning techniques, their application to tax arrears prediction remains underexplored. We forecast individual local tax defaults in Republic of Korea using machine and deep learning algorithms, including convolutional neural networks (CNN), long short-term memory (LSTM), and sequence-to-sequence (seq2seq). Our model incorporates diverse credit and public information like loan history, delinquency records, credit card usage, and public taxation data, offering richer insights than prior studies. The results highlight the superior predictive accuracy of the CNN model. Anticipating local tax arrears more effectively could lead to efficient allocation of administrative resources. By leveraging advanced machine learning, this research offers a promising avenue for refining tax collection strategies and resource management.

Keywords: Deep learning, financial machine learning, local tax delinquency, tax defaulter prediction, tax information systems.

This study is based on the following doctorate dissertation by the corresponding author, Eunchan Kim:
Kim, E. (2023). *Exploring the Application Studies of Artificial Intelligence for Tabular and Image Data: Insights from Radiology, Taxation, and Security Detection Research Fields.* Seoul National University.

http://doi.org/10.3837/tiis.2024.01.003 **ISSN : 1976-7277**

1. Introduction

In 2016, a significant policy shift known as "Government 3.0" was introduced in the Republic of Korea. This transformative initiative aimed to incorporate scientific principles into public administration, and as part of this endeavor, the Ministry of Public Administration and Security began exploring the implementation of scientific tools, including a standardized analysis model for public big data [1]. From its inception, Government 3.0 emphasized a data-driven approach to enhance scientific public administration [2]. The year 2020 saw the enactment of a new law [3], further accelerating the adoption of this data-centric approach through mandatory sharing of large datasets among public institutions. This legal framework not only mandated data sharing but also established a data integration management platform to facilitate efficient data provisioning, linkage, and collaborative utilization. Consequently, the government's inclination to address unresolved challenges using data-driven analyses has surged, prompting various local governments to actively seek solutions through data utilization.

Within this context, extensive efforts have been directed towards applying scientific principles to address the persistent issue of local tax arrears, a concern of considerable magnitude for local administrations nationwide [1, 4]. Notable progress has been achieved in estimating individual local tax arrears using logistic models, along with the strategic prioritization of data collection based on model predictions. Nevertheless, prior studies are limited by their reliance solely on simple logistic regression or shallow machine learning models to compute probability values. Deep learning, which has recently appeared, is showing excellent performance in various fields such as image processing, natural language processing, item recommendation and speech recognition [5-10]. There are attempts to borrow deep learning in the financial field as well and this paper presents a method of using deep learning models to predict local tax delinquents.

This study introduces a more comprehensive model for predicting local tax arrearages. First, by incorporating both public and credit information, the research process and results are significantly refined. Second, a nation-wide stratified random sampling of individual local tax defaulters is employed to ensure fair prediction of default instances across the entire country. Last, by leveraging diverse machine-learning algorithms, this study proposes the utilization of distinct scientific administrative processing methods, all within the framework of an artificial intelligence (AI)-based approach.

2. Related work

2.1 Local Tax Defaulters

Since local taxes are a major financial resource for local governments, analysis of tax collection and management is very important and it is also crucial for local governments to establish appropriate local tax policies [11, 12]. Based on the 2022 local tax settlement data provided by the Ministry of the Interior and Safety, the accumulated local tax arrears for 2021 stood at approximately 3.4 trillion won. This figure has exhibited a consistent upward trend in recent years [13]. As demonstrated in **Table 1**, the majority of default cases and the highest quantum of tax arrears are concentrated within the capital region. Notably, Seoul and Gyeonggi Provinces account for approximately 41.78% of defaulting entities and around 52.31% of the overall sum of local tax arrears as shown in **Fig. 1**. Given that local taxes constitute a pivotal revenue stream for regional administrations, the escalating volume of outstanding local taxes underscores the imperative of robust scientific public administration measures to address this fiscal exigency effectively.

In light of the cardinal role local taxes fulfill in furnishing the financial substratum for regional administrations, the burgeoning volume of unresolved local tax arrears assumes an urgency that cannot be understated [14]. This landscape underscores the exigency of formulating and implementing robust measures within the realm of scientific public administration. An administrative strategy woven from empirical evidence and delicate foresight, these measures hold the potential to not merely rectify fiscal imbalances but to foster a sustainable paradigm for financial growth within the administrative sphere.

Table 1. Number of local tax defaulters and the amount of local taxes in arrears for each province of the Republic of Korea in 2021.

Province	**No. of local tax defaulters (in thousands)**	**Amount of local tax arrears (in billions of won)**
Total	6,584	3,397.9
Seoul	864	746.6
Busan	432	174.5
Daegu	301	72.6
Incheon	465	193.6
Gwangju	140	52
Daegeon	170	61.7
Ulsan	125	60.5
Sejong	38	22.7
Gyonggi	1,887	1,030.7
Gangwon	210	77.4
Chungbuk	214	89.8
Chungnam	303	147.6
Jeonbuk	245	93.5
Jeonnam	244	79.1
Gyeongbuk	388	179
Gyeongnam	465	234.8
Jeju	93	81.7

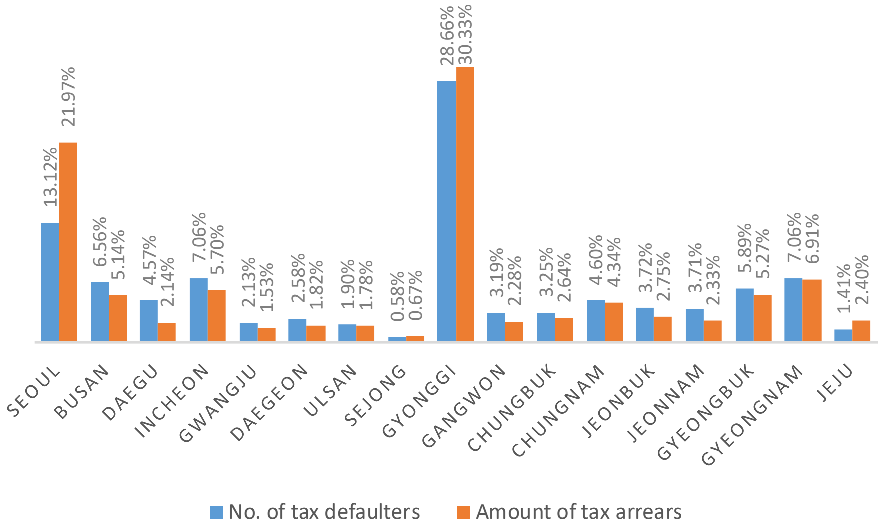

Fig. 1. Percentages in number of local tax defaulters and the amount of local taxes in arrears for each province of the Republic of Korea in 2021.

In the context of the Republic of Korea, individuals are granted a designated period of 30 days subsequent to the imposition of personal local taxes within which to effectuate their initial payment. Should this initial deadline transpire without payment, an individual is automatically designated as a defaulter within the tax framework. Subsequent to this classification, local tax authorities proceed to recover overdue local taxes through a process involving the exertion of pressure upon defaulting individuals or the lawful seizure of their assets [4].

To elaborate, should an individual fail to remit payment within 30 days following the issuance of a local tax notification, they are formally categorized as a defaulter. The spectrum of local taxes, which are levied and collected annually to meet the fiscal requisites of local governance entities, encompasses a diverse array of 11 categories, encompassing taxes related to automobiles, residences, property acquisitions, property ownership, local income, registrations and licenses, tobacco consumption, leisure activities, regional development initiatives, local consumption, and local educational undertakings [15]. Within this array, a sum total of six distinct tax components possesses the potential to accrue arrears, encompassing taxes associated with registrations and licenses, automobiles, residences, local income, and property acquisitions [15, 16].

As previously alluded to, the prognostic ability to anticipate instances of local tax default not only stands to fortify the fiscal robustness of local governing bodies, but also to facilitate the streamlined collection of local taxes. Once a 50-day default threshold has been crossed, supplementary administrative measures, including the issuance of notifications and other coercive mechanisms, are set into motion. Therefore, the realm of efficient administrative oversight stands feasible, contingent upon our capacity to preemptively identify and forecast those individuals most predisposed to default on their local tax obligations.

2.2 Tax and Loan Default Prediction Using Machine Learning

The anticipation of default risk stands as a prominent and extensively investigated concern within the domains of business, finance, and economics. A plethora of methodologies grounded in machine learning have been posited for the evaluation of credit risk [17–20]. Li et al. [21] have noted that XGBoost [22] exhibits superior predictive capabilities in relation to credit risk, surpassing the performance of logistic regression while displaying heightened stability. Further, Kruppa et al. [23] have observed that a random forest [24] demonstrates superior performance compared to finely-tuned logistic regression and k-nearest neighbors (k-NN).

Beyond the realm of credit risk, contemporary research has ventured into the prediction of tax and loan delinquencies. Abedin et al. [25] have demonstrated the efficacy of feature transformations, encompassing logarithmic and square-root transformations, in enhancing the efficacy of existing statistical methodologies and machine learning techniques geared towards the prognostication of local tax arrears. However, it should be noted that their study differs from ours in terms of the machine learning models employed; they predominantly employed simpler algorithms such as k-NN, linear discriminant analysis (LDA) [26], and decision trees, as opposed to delving into more intricate deep learning paradigms like CNN [27] and LSTM networks [28]. Höglund [29] has showcased noteworthy performance in the prediction of tax defaults via a genetic-algorithm-driven variable selection approach. Huang et al. [30] have prognosticated financial distress within corporate entities through a medley of supervised and unsupervised learning algorithms. Nevertheless, it is worth noting the constraints of their study, which encompassed a modest sample size of 64 companies and a total of 16 financial indicator features.

In parallel with the pursuit of tax default prediction, certain inquiries have steered machine learning methodologies towards the projection of mortgage loan delinquencies. Noteworthy instances include [31] and [32], where authors harnessed elementary machine learning tools to ascertain the likelihood of mortgage loan delinquency. Furthermore, [33] presents a thorough parametric comparative analysis of diverse machine learning models for the prediction of mortgage loan delinquencies, albeit with relatively limited utilization of advanced deep learning techniques such as CNNs and attention-equipped seq2seq models. Abedin et al. [34] have expounded upon the application of machine learning to identify delinquent corporate taxpayers. The authors accentuate the potency of logarithmic transformations among various feature manipulation methods and identify random forests as the most efficacious machine learning model among their applied arsenal. Similarly, Gebauer et al. [35] have embarked on an exploration of the utility of value-added tax (VAT) evasion as a predictive indicator for corporate tax delinquency status.

3. Local Tax Defaulter Prediction Methods

In this research, we looked at a lot of different machine learning methods that can predict. We organized these methods into three groups and used them in a specific way. First, we used something called the seq2seq algorithm, which is usually used to help with translations between languages. It's like a tool that understands patterns in data over time, which can be really helpful for data that changes over time, like stock prices. Then, we used two more methods. One is called a CNN, which is good at understanding information from nearby data points. It's like looking at the neighborhood around a data point to understand it better. The other method is a ResNet, which is like an improved version of the CNN. It combines the information from different layers to get a better picture of the data. Lastly, we used two other methods to make predictions about people who might not pay their taxes on time. These methods are the random forest and the support vector machine (SVM). They're like popular tools that can help us make predictions about things without going too deep into the complex details. To sum up, we tried out six different methods to predict whether something might go wrong, like someone not paying taxes on time. A brief overview of the six models used for tax defaulter prediction is as follows.

3.1 LSTM

LSTM is an extended model of the RNN algorithm, and it can remember older information more effectively than an RNN owing to its long-term memory [28]. Because RNNs by nature learn using their short-term memory, there is a long-term dependency problem in which the information that comes out during the early phase vanishes during the later phase. The cell state of the LSTM solves this problem. The cell state, a key aspect of LSTM, has several elements called ‘gates’ that allow eliminating unnecessary information or adding and storing new information. It also acts as a gradient highway to prevent a vanishing gradient, allowing information to spread farther. LSTM consists of four stages. As the first step, LSTM uses a forget gate to determine how much previous information has been forgotten.

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{1}$$

The forget gate ($\mathrm{f_t}$) takes the previous hidden state ($\mathrm{h_{t-1}}$) and current input data ($\mathrm{x_t}$) as inputs and converts them into values between zero and 1 as they pass through the sigmoid function. If the output that passes through the forget gate has a value of zero, the past memory completely disappears. Conversely, if the output passing through the forget gate has a value of 1, the past memory is completely preserved. The second step is the input gate ($\mathrm{i_t}$), which determines how much of the current data will be reflected. As with the forget gate, it takes $\mathrm{h_{t-1}}$ and $\mathrm{x_t}$ as inputs, which are converted into values between zero and 1 using the sigmoid function. We also obtained the new information, $\tilde{C}_t$, using $\mathrm{h_{t-1}}$ and $\mathrm{x_t}$:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i \tag{2}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{3}$$

In other words, we compute new information ($\tilde{C}_t$) to be stored in the cell state and how much new information (i_t) has been stored. During the third step, we update the cell state. After determining the amount of past information to forget using $\mathrm{f_t}$, and how much current information is to be reflected using $\mathrm{i_t}$, their sum becomes the input to the next cell state as follows:

$$C_t = f_t \circ C_{t-1} + i_t \circ \tilde{C}_t. \tag{4}$$

The final step is to determine the value of $\mathrm{o_t}$ to be sent to the output. We multiply $\mathrm{o_t}$ by tanh ($\mathrm{C_t}$) to determine $\mathrm{h_t}$, which will be entered as the input of the next state.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o \tag{5}$$

$$h_t = o_t \circ \tanh(C_t) \tag{6}$$

3.2 Seq2seq with Attention

Seq2seq [36], which is widely used in machine translation and document summaries, receives a sequence of words as input, and then outputs the sequence of words. Seq2seq consists of an encoder and decoder, and an RNN is primarily used. In the case of machine translation, the encoder generates the context of one sentence as a vector, which is passed to the decoder. For example, in an English–Korean translation, the encoder encodes one English sentence sequence into one context vector (compression). The decoder then decodes the encoded context vector to generate the corresponding Korean sentence (reproduction).

However, because this seq2seq model represents the context of an entire sentence as a single fixed vector, it has difficulty handling particularly long sentences. The longer the sentence is, the more information that disappears regarding the words that appeared first. Bahdanau et al. [37] introduced an attention technique to solve this problem. The seq2seq model with attention solves the aforementioned problem by setting an attention score, which is a weight, for each word as the input to the model. For all words of both the encoder and the decoder, an attention score is calculated through a specific operation (e.g., dot product), which is reflected in the output of the decoder. This seq2seq model performs well in language-related domains and is also known to a show good classification performance. Therefore, this model was selected as a candidate.

3.3 CNN

A CNN is a model that mimics the human optic nervous system and since the advent of AlexNet has been a leader in the deep-learning era [27]. A CNN consists of a convolution layer and a pooling layer, and achieves a good performance when the locality of the data is important within the spatial structure. In particular, it shows an excellent performance in image processing, where the spatial structure and locality are extremely important. Because the convolution layer only connects to the input data or the output of the previous layer belonging to the received field, only the part belonging to that received field reflects the information. The pooling layer extracts only the necessary information from the window. Typical examples are max pooling, which extracts only the maximum value from the window, and average pooling, which extracts the average value of the window. If we stack more convolution and pooling layers, we can obtain information regarding a larger area in an image or video.

3.4 ResNet

Deep learning techniques, including a CNN, perform well in various fields; however, when the layer is too deep, the performance deteriorates owing to a vanishing gradient or exploding problem. Most artificial neural network models update the weights using a backpropagation method after determining the derivative of the loss function with respect to each weight. The deeper the neural network is, the more derivatives that are multiplied. If a small derivative is multiplied several times, it approaches zero, which indicates a gradient loss. Conversely, when a large derivative is multiplied several times, the value becomes extremely large, which is called a gradient explosion. In general, artificial neural networks are predicted to perform better as the layers deepen. However, in [38], the authors showed that a plain 20-layer network has fewer training and test errors than a plain 50-layer network for an image classification problem, which they insist is not caused by an overfitting. This problem is called degradation and is caused by the loss of gradients. To solve this problem, ResNet, which uses the skip connection method in which input x is added to the output value after several layers, has been proposed. Let us assume that when the input value of a neural network is x, the value created through several neural-network layers is F(x). In neural networks, the goal is to find a function H(x) that maps an input x to a target y, typically when H(x) = F(x). However, ResNet, which uses a skip connection, defines H(x) = F(x) + x and optimizes it. Using ResNet, the authors built deeper neural networks and achieved a better performance.

3.5 Random Forest

A single decision tree is extremely susceptible to an overfitting, and a random forest using multiple decision trees is applied as an ensemble approach to solve the overfitting problem. A random forest uses a process called bagging, which applies selected subsets of data or features, rather than using all data and features, to build each tree [24]. When we model a decision tree, we consider all features, select those with the highest information gain, and split the data based on such gain. However, with a random forest, we create multiple trees, where each tree considers only a fraction of the total features. Here, the term ‘random’ means that when creating each decision tree, the features to be used are randomly selected. After creating multiple trees, voting is conducted on the results of the classification of each tree, and the final result is selected by a majority vote. A random forest is excellent at preventing an overfitting and improves the prediction accuracy. In addition, for classification tasks, we can rank relatively important features.

3.6 Support Vector Machine

An SVM is a model that defines a decision boundary, which is a criterion for classification [39, 40]. Classification is applied on new data by determining which side of the boundary it falls on. This decision boundary is a line in two dimensions, a plane in three dimensions, and a hyperplane in higher dimensions. Support vectors are the closest data points to the decision boundary. The goal of a support vector machine is to find a decision boundary that maximizes the distance between the support vectors. The margin is the distance between the decision boundary and the support vectors. In other words, with an SVM, the optimal decision boundary maximizes the margin. If the given data are linearly separable, the decision boundary can be found directly through an optimization, which minimizes the objective function L.

$$\begin{gathered} L = \min \frac{1}{2} \|W\|^2, \\ s.t \;\; y_i(w \cdot x_i + b) \geq 1, \end{gathered} \tag{7}$$

where y_i is the label, and x_i is the input data. Therefore, SVM can be expressed as follows for linearly separable data:

$$f(x) = w^t x + b = \sum_{i=1}^{N} y_i \alpha_i x_i x + b. \tag{8}$$

However, the above formula does not allow any errors. If the data are not divided by a straight boundary line and a linear separation is impossible, we can allow errors using the slack variable ξ, as shown below. This means finding the hyperplane while allowing some errors to occur. Here, C is a hyperparameter that determines the strength of the error penalty.

$$\begin{gathered} \min \frac{1}{2} \|W\|^2 + C \sum_{i=1}^{N} \xi_i \\ s.t \;\; y_i(w \cdot x_i + b) \geq 1 \end{gathered} \tag{9}$$

Converting the above formula into a dual problem using a Lagrangian is as follows.

$$\begin{gathered} \min \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j (x_i \cdot x_j) - \sum_{i=1}^{N} \alpha_i \\ s.t. \sum_{i=1}^{N} y_i \alpha_i = 0, a \leq \alpha_i \leq C \end{gathered} \tag{10}$$

In addition, if the data are not linearly separable, a kernel trick is used to map the data into a high-dimensional space and then determine the decision boundaries in that hyperplane. Instead of using the original D-dimensional independent variable vector x, we can use the M-dimensional vector $\phi(x)$ transformed by the basis function as the independent variable. ($\phi(\cdot): R^D \rightarrow R^M$) In this case, the original vector x is transformed into

$$x = (x_1, x_2, \dots x_D) \rightarrow \phi(x) = (\,\phi_1(x), \phi_2(x), \dots \phi_M(x)\,). \tag{11}$$

For the above expression, we change x into $\phi(x)$ with a basis function transformation, and because all basis functions are used only in the form of $\phi(x_i)^T\phi(x_j)$, they can be represented as $K(x_i, x_j)$. The resulting formula is as follows:

$$\min \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j K(x_i \cdot x_j) - \sum_{i=1}^{N} \alpha_i, \tag{12}$$

$$s.t. \sum_{i=1}^{N} y_i \alpha_i = 0, a \leq \alpha_i \leq C.$$

Kernel $K(x_i, x_j)$ can be viewed as a criterion to measure the similarity between two data samples, and the linear kernel and radial basis function (RBF) kernel are often used.

4. Experiment

4.1 Dataset

In this research, we used data on both credit and public information about taxpayer (including people who did not pay their taxes on time, as well as those who paid their taxes regularly) in the Republic of Korea. We collected this data over three fiscal years, from January 2015 to December 2017. We used the same set of data that of the previous work [1] has used before. To achieve a well-rounded sample, we employed random selection to choose individuals nationwide. This encompassed both punctual local tax payers and defaulters.

Further, by using a stratified random sampling method, we factored in the regional distribution outlined in **Fig. 1**. As a result, we gathered a dataset of 354,130 individuals (excluding companies) for analysis maintaining the same distribution by region. We looked at about 190 different pieces of information about each person, removed many of the variables that provided excessive information about the tax delinquent, and settled on about 50 final variables. Those selected variables we used is listed in **Table 2**. The main thing we were trying to find out is whether someone paid their local taxes on time or not. If someone did not pay their taxes within 30 days of getting a tax notice, they were considered a tax defaulter. This can be checked automatically by the tax information system.

Out of all the individuals we studied, about 12.05% did not pay their taxes on time, and in our sample, it was about 12.07%. These numbers are very close, which tells us that the data we have is very trustworthy and reliable.

Table 2. Selected variables: The target variable is the status of local tax arrears within public arrears information, while other variables serve as training for the target variable.

Credit information	Loan/ overdue information	the number of lenders, number of loans, loan amounts, remaining loan balance, loan expiration date, loan status, instances of overdue payments, status of overdue cases, instances of canceled defaults, canceled overdue payments, days after cancellation of delinquencies, credit scores
	Credit card information	the amount spent using credit cards, the number of credit cards, amounts spent using debit and credit cards, credit limits, lump-sum and installment spending amounts, installment usage rate, short-term card loan amounts, counts of short-term card loans, long-term card loan amounts, counts of long-term card loans, the number of card companies used, loan expiration dates, loan balances, cases of overdue payments, overdue status, instances of canceled defaults, canceled overdue payments, days after cancellation of delinquencies
Public information	Tax information	the time since local taxation, the amount of local taxes, types of local taxes, the status of local tax arrears (which is the target variable), the number of local taxes, amounts of local taxes, the government's urgency regarding local tax payment.
	Public information	missing persons, individuals declared as incompetent or quasi-incompetent, those undergoing personal bankruptcy or rehabilitation, individuals involved in disclaimers or lotteries, cases of transaction suspension, deaths, suspicions linked to multiple resident numbers, repayment capabilities, annual income, mobile communication patterns, utility bill payment history, employment status, employment duration, statuses of major insurances (social insurance, national pension, health insurance, and employment insurance), durations of being part of these insurances, occupation details, information on individuals wanted by Interpol and the North Korean defectors

To mitigate the influence of COVID-19 and ensure a reliable assessment process, we designated the fiscal period spanning from January 2015 to December 2017. Within the dataset encompassing 354,130 individual local taxpayers (excluding corporations), we performed a division, creating two distinct subsets: a training set covering the initial two fiscal years, and a test set encompassing the subsequent fiscal year. This division facilitated a comprehensive evaluation of the model's efficacy and contributed to enhancing its overall precision. The dataset corresponding to the latter fiscal year was earmarked for test and validation purposes.

4.2 Computing resources

We used an Nvidia RTX 3080 GPU, 4 Intel i7-10700 CPUs, and 128 GB of RAM for the experiments. The analysis was conducted using R version 4.2.1.

4.3 Architectures of each model

In this study, we assessed performance using a measure called accuracy. As shown in **Table 3**, accuracy is calculated for two groups: local tax defaulters and non-defaulters (referred to as "normal"). The accuracy formula is ((TP + TN)/(TP + FP + TN + FN)), and the weighted average considers the size of each group. Here, TP stands for true positive, indicating instances where an actual non-default is correctly predicted as a non-default. FP represents false positive, which occurs when an actual default is incorrectly predicted as a non-default. Similarly, FN signifies false negative, when an actual non-default is inaccurately predicted as a default. Lastly, TN stands for true negative, indicating situations where an actual default is correctly predicted as a non-default.

Table 3. Confusion matrix for local tax default prediction.

Actual **Prediction**	**Non-default**	**Default**
Non-default	TP (True Positive)	FP (False Positive)
Default	FN (False Negative)	TN (True Negative)

The structure of each predictive model is presented in Table 4. Here are the detailed specifications: The LSTM model was crafted using five layers arranged as (512 × 256 × 128 × 32 × 2), while the seq2seq model consisted of four layers configured as (256 × 128 × 32 × 2). In the case of the CNN model, it involved two convolution layers followed by a pooling layer (conv → conv → pooling). This was further accompanied by five fully connected layers structured as (2024 × 512 × 128 × 64 × 2). The ResNet model featured four layers (ResNetLayer → Pooling), succeeded by three fully connected layers (512 × 128 × 2). These architectural choices were intentionally simplified to ensure practical implementation and operational utility of the models, while simultaneously minimizing the necessity for extensive parameter adjustments.

Table 4. Architecture and hyper parameters of each model.

Model	**Architecture & Hyper Parameters**
SVM	Kernel = RBF
Random Forest	Estimators =20, Leaf Node=20, gini
LSTM	5 layers(512, 256, 128, 32, 2)
ResNet	(ResNetLayer→pooling)×4+3 layers
Seq2Seq with attention	4 layers(256, 128, 32, 2) with attention
CNN	(conv→conv→pooling)×3+5 layers

4.4 Experimental results

For the purpose of comparison with prior research endeavors, the outcomes of earlier investigations - specifically, the accuracies of the Logistic Regression and Random Forest models - were integrated as baseline results, a portrayal presented in **Table 5**. While acknowledging the disparity in the modeling scope between the current and previous studies - where the previous study solely utilized credit information - it remains challenging to effectuate a direct comparison. Nevertheless, a discernible effect of data convergence is evident through the showcase of approximately 4% divergence in the application of the identical Random Forest algorithm.

Exploring the performance evaluation of the models utilized in this study to predict local tax arrears, significant patterns come to light. The CNN model takes the lead, achieving the highest accuracy, followed by seq2seq with an attention mechanism. Concurrently, other models exhibit comparable performance, showcasing accuracies within the range of 82% to 83%. Drawing from the preceding research outcomes [1], our models' predictive prowess emerges as remarkably elevated. Importantly, the construction of relatively streamlined models in this study corresponds with practical administrative considerations. Consequently, it's conceivable that our ResNet model, featuring a less complex structure, may not have reached its optimal performance zenith. Thus, informed by this context and the broader administrative context, we have concluded that our meticulously developed CNN model best serves as our ultimate choice for forecasting tax delinquencies.

Table 5. Various model prediction accuracies for local tax default prediction.

Model	**Accuracy (%)**
Baseline 1: Logistic Regression from the Previous Study	78.00
Baseline 2: Random Forest from the Previous Study	78.86
SVM	82.23
Random Forest	82.86
LSTM	83.39
ResNet	83.72
Seq2Seq with attention	84.43
CNN	**87.64**

Based on the accuracy results of this study and in consideration of the preceding research, our straightforward CNN model exhibited the highest accuracy. Notably, when inspecting the result's confusion matrix as shown in **Table 6**, an intriguing observation emerged: there were no instances wherein an individual who had met their tax obligations was erroneously labeled as a delinquent. This particular aspect motivated our selection of the CNN model as the definitive choice. This decision aligns with the overarching administrative objective of effective prediction, emphasizing the importance of accurately distinguishing between those who are tax compliant and those who might be perceived as delinquent. This distinction is pivotal, as misclassifying tax-compliant individuals as delinquents could introduce critical administrative complications and predicaments. Hence, the prudent selection of the CNN model is a judicious step towards preventing such issues, drawing upon insights garnered from our prior explorations.

Table 6. Confusion matrix of our CNN model.

Actual **Prediction**	**Non-default**	**Default**
Non-default	295,176	45,011
Default	**0**	23,853

To further enrich our understanding of the comparative performance of the various models and to establish their robustness, a comprehensive assessment was conducted using the area under the receiver operating characteristic (AUROC) curve. This analytical approach, visualized in **Fig. 2**, represents an extension of our previous exploration, allowing us to delve deeper into the intricacies of the models' predictive capabilities. Notably, the AUROC curve serves as a potent tool for evaluating the models' ability to discriminate between positive and negative instances across a spectrum of decision thresholds.

The results of this AUROC analysis substantiate the consistency and stability of the models' performance. Each model consistently maintained a steady trajectory across various ROC levels, indicating their reliability and steadfastness in delivering predictive insights. It is important to highlight that the AUROC graph serves as a valuable supplementary method for confirming the predictive efficacy of models, a validation step that complements our previous findings. Crucially, our investigation revealed a compelling aspect: there were no significant differentiations discernible among the models' AUROC curves. This outcome underscores the equitable performance of the models and reiterates their comparable predictive abilities. The convergence of these results reinforces the validity and coherence of our earlier work, affirming the models' consistent and dependable performance levels across various evaluation dimensions.

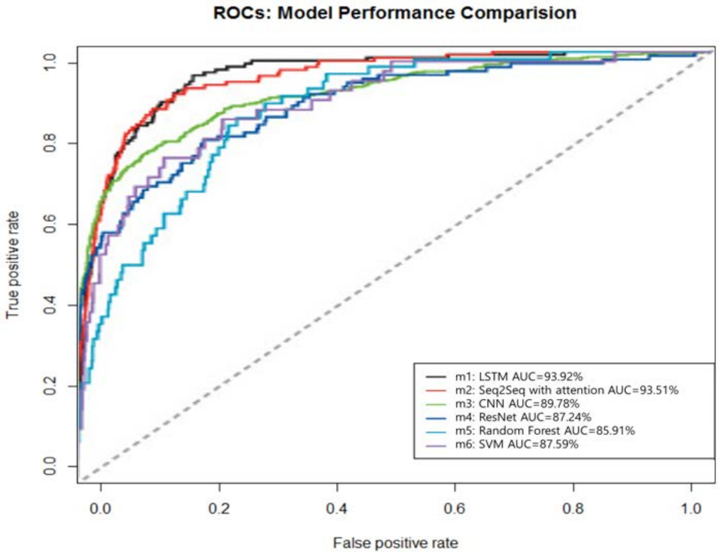

Fig. 2. AUROC curves for different models in prediction of local tax defaulter status.

5. Conclusion

This research delved into the realm of local tax arrears prediction, employing a diverse array of machine learning models. What sets this study apart is its innovative approach that harnesses both public and credit information, distinguishing itself from previous investigations that predominantly relied on either public or financial data. In a departure from convention, our methodology incorporates a stratified random sampling technique, enabling the development of predictive models that offer comprehensive coverage across the entire nation. Furthermore, while previous studies on tax defaulters used regression and shallow machine learning models, this study embraces an eclectic suite of deep and machine learning models, in line with contemporary trends in the domains of data science and artificial intelligence. Within this diversified landscape, our CNN-based methodology emerged as a standout performer, showcasing an impressive accuracy of 87.64%. It is pertinent to note that the other models, while varying in structure, consistently demonstrated commendable performance, each boasting a minimum accuracy of 82%. Looking forward, our research trajectory encompasses the incorporation of elucidatory techniques such as the shapley value (SHAP) and Lime. These techniques hold the promise of augmenting not only the predictive accuracy of our proposed methodology but also its interpretability. As we navigate the course of future research, these enhancements aim to further elevate the effectiveness and transparency of our predictive framework, aligning with the progressive evolution of data science and AI practices.

Acknowledgement

This study was supported by the Institute of Management Research at Seoul National University.

References

[1] E. Kim, and B. Yoo, “Credit information and public big data analysis: Development of prediction model for the possibility of recovering tax arrears and improvement of tax arrears information system,” in *Proc. of International Conference on Korea Society of Management Information System*, pp. 6-12, May. 2018. Article (CrossRef Link)

[2] E. Kim, M. Kim, and Y. Kyung, “A case study of digital transformation: Focusing on the financial sector in South Korea and overseas,” *Asia Pacific Journal of Information Systems*, vol. 32, no.3, pp.537-563, 2022. Article (CrossRef Link)

[3] Act on Revitalization of Data-Based Administration (abbreviation: Data-Based Administration Act) [Enforcement 2020. 12. 10.] [Act No. 17370, 2020. 6. 9., enacted] Ministry of Public Administration and Security (Public Data Policy Division)

[4] H. Byun, “Securitization of local tax receivables and liens - A solution to the insufficiency of local tax revenue,” *Seoul Law Review*, vol. 26, no. 2, pp. 413-454, 2018. Article (CrossRef Link)

[5] R. Mu, and X. Zeng, “A review of deep learning research,” *KSII Transactions on Internet and Information Systems*, vol. 13, no. 4, pp.1738-1764, 2019. Article (CrossRef Link)

[6] H. Ma, and Q. Liu, “In-depth recommendation model based on self-attention factorization,” *KSII Transactions on Internet & Information Systems*, vol. 17, no. 3, pp. 721-739, 2023. Article (CrossRef Link)

[7] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521 no. 7553, pp. 436-444, 2015. Article (CrossRef Link)

[8] A. Shrestha, and A. Mahmood, “Review of deep learning algorithms and architectures,” *IEEE Access*, vol. 7, pp. 53040-53065, 2019. Article (CrossRef Link)

[9] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, “Deep learning for computer vision: A brief review,” *Computational Intelligence and Neuroscience*, vol. 2018, pp. 1-13, 2018. Article (CrossRef Link)

[10] V. Sorin, Y. Barash, E. Konen, and E. Klang, “Deep learning for natural language processing in radiology—fundamentals and a systematic review,” *Journal of the American College of Radiology*, vol. 17, no. 5, pp. 639-648, 2020. Article (CrossRef Link)

[11] J. Alm, R. D. Buschman, and D. L. Sjoquist, “Rethinking local government reliance on the property tax,” *Regional Science and Urban Economics*, vol. 41, no. 4, pp. 320-331, 2011. Article (CrossRef Link)

[12] I. Belmonte-Martin, L. Ortiz, and C. Polo, “Local tax management in Spain: A study of the conditional efficiency of provincial tax agencies,” *Socio-Economic Planning Sciences*, vol. 78, pp. 101057, 2021. Article (CrossRef Link)

[13] Ministry of Interior and Safety, “2022 Local Tax Statistical Yearbook (Fiscal Year of 2021),” [Online]. Available:

https://www.mois.go.kr/frt/bbs/type001/commonSelectBoardArticle.do;jsessionid=2kNKa4-VerVGnD0Bn+U8nXGu.node50?bbsId=BBSMSTR_000000000014&nttId=96011

[14] H. J. Smith, S. Park, and L. Liu, “Hardening budget constraints: a cross-national study of fiscal sustainability and subnational debt,” *International Journal of Public Administration*, vol. 42, no. 12, pp. 1055-1067, 2019. Article (CrossRef Link)

[15] Y. Ryoo, “A study on the improvement method of the local tax bill delivery service using the information system,” *Journal of Korean Association for Regional Information Society*, vol. 22, no. 1, pp. 53-73, 2019. Article (CrossRef Link)

[16] J. J. Moon and K. Hong, “A study on the tax source adjustment plan of national and local taxes for refining local finance,” *Journal of Tax Studies*, vol. 20, no. 4, pp. 179-200, 2020. Article (CrossRef Link)

[17] G. Wang, J. Hao, J. Ma, and H. Jiang, “A comparative assessment of ensemble learning for credit scoring,” *Expert Systems with Applications*, vol. 38, no. 1, pp. 223-230, Jan. 2011. Article (CrossRef Link)

[18] L. Yu, S. Wang, and K. K. Lai, “Credit risk assessment with a multistage neural network ensemble learning approach,” *Expert Systems with Applications*, vol. 34, no. 2, pp. 1434-1444, Feb. 2008. Article (CrossRef Link)

[19] S. Lessmann, B. Baesens, H. V. Seow, and L. C. Thomas, “Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research,” *European Journal of Operational Research*, vol. 247, no. 1, pp. 124-136, Nov. 2015. Article (CrossRef Link)

[20] S. Bhatore, L. Mohan, and Y. R. Reddy, “Machine learning techniques for credit risk evaluation: A systematic literature review,” *Journal of Banking and Financial Technology*, vol. 4, no. 1, pp. 111-138, June. 2020. Article (CrossRef Link)

[21] Y, Li, “Credit risk prediction based on machine learning methods,” in *Proc. of the 14th International Conference on Computer Science & Education*, Toronto, Canada, pp. 1011-1013, Aug. 2019. Article (CrossRef Link)

[22] T. Chen, and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, USA, pp. 785-794, Aug. 2016. Article (CrossRef Link)

[23] J. Kruppa, A. Schwarz, G. Arminger, and A. Ziegler, “Consumer credit risk: Individual probability estimates using machine learning,” *Expert Systems with Applications*, vol. 40, no. 13, pp. 5125-5131, 2013. Article (CrossRef Link)

[24] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh, “Investigation of the random forest framework for classification of hyperspectral data,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 3, pp. 492-501. Mar. 2005. Article (CrossRef Link)

[25] M. Z. Abedin, G. Chi, M. M. Uddin, M. S. Satu, M. I. Khan, and P. Hajek, “Tax default prediction using feature transformation-based machine learning,” *IEEE Access*, vol. 9, pp. 19864-19881, Dec. 2020. Article (CrossRef Link)

[26] S. Balakrishnama and A. Ganapathiraju, “Linear discriminant analysis for signal processing problems,” in *Proc. of IEEE Southeastcon’99*, pp. 78-81, 1999. Article (CrossRef Link)

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84-90, June. 2017. Article (CrossRef Link)

[28] S. Hochreiter, and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997. Article (CrossRef Link)

[29] H. Höglund, “Tax payment default prediction using genetic algorithm-based variable selection,” *Expert Systems with Applications*, vol. 88, pp. 368-375, Dec. 2017. Article (CrossRef Link)

[30] Y. P. Huang, and M. F. Yen, “A new perspective of performance comparison among machine learning algorithms for financial distress prediction,” *Applied Soft Computing*, vol. 83, pp. 105663, Oct, 2019. Article(CrossRef Link)

[31] S. Chen, Z. Guo, and X. Zhao, “Predicting mortgage early delinquency with machine learning methods,” *European Journal of Operational Research*, vol. 290, no. 1, pp. 358-372, Apr. 2021. Article (CrossRef Link)

[32] L. Lai, “Loan default prediction with machine learning techniques,” in *Proc. of 2020 International Conference on Computer Communication and Network Security (CCNS)*, Xi'an, China, pp. 5-9, Aug. 2020. Article (CrossRef Link)

[33] S. E. A. Ali, S. S. H. Rizvi, F.W. Lai, R. F. Ali, and A. A. Jan, “Predicting delinquency on Mortgage loans: An exhaustive parametric comparison of machine learning techniques,” *International Journal of Industrial Engineering and Management*, vol. 12, no. 1, pp. 1-13, Mar. 2021. Article (CrossRef Link)

[34] M. Z. Abedin, M. K. Hassan, I. Khan, and I. F. Julio, “Feature transformation for corporate tax default prediction: Application of machine learning approaches,” *Asia-Pacific Journal of Operational Research*, vol. 39, no. 4, pp. 2140017, 2022. Article (CrossRef Link)

[35] A. Gebauer, C. W. Nam, and R. Parsche, “Can reform models of value added taxation stop the VAT evasion and revenue shortfalls in the EU?,” *Journal of Economic Policy Reform*, vol. 10, no. 1, pp. 1-13, 2007. Article (CrossRef Link)

[36] I. Sutskever, O. Vinyals, and Q.V. Le, “Sequence to sequence learning with neural networks,” in *Proc. of the 27th Conference on Neural Information Processing Systems*, vol. 2, pp. 3104-3112, Dec. 2014. Article (CrossRef Link)

[37] D. Bahdanau, K. Cho, K, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *Proc. of International Conference on Learning Representations*, San Diego, CA, USA, May. 2015. Article (CrossRef Link)

[38] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, June. 2016. Article (CrossRef Link)

[39] W. S. Noble, “What is a support vector machine?,” *Nature Biotechnology*, vol. 24, no. 12, pp. 1565-1567, Dec. 2006. Article (CrossRef Link)

[40] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18-28, 1998. Article (CrossRef Link)

YongHyun Lee received the B.A. degree in computer engineering from the Sungkyunkwan University in 2015. He is currently pursuing his Ph.D. degree at the Department of Computer Science and Engineering, Seoul National University. He is working as a researcher at both Seoul National University Hospital and Jeonbuk National University Hospital. His research interests include artificial intelligence, data science, graph neural network, graph classification, medical and financial data mining.

Eunchan Kim received the B.A. degree in economics from the University of Minnesota, Twin Cities in 2012, and the M.S. degree in management (concentration: information systems) and the Ph.D. degree in engineering (concentration: intelligence and information) from Seoul National University in 2017 and 2023, respectively. He is currently working as a senior researcher at Hanwha Group, and a lecturer at the College of Business Administration, Seoul National University. He also serves as a visiting scholar at both Seoul National University Hospital and Jeonbuk National University Hospital. His research interests include information systems, artificial intelligence (AI), applications of AI, and engineering management.