

¹Yubin Wang*

Prediction of Student Employability through Internship based on Big Data Analysis



Abstract: - The internship course is among the most important because it gives students a hands-on opportunity to apply their knowledge and get ready to launch a professional career. Internships, yet, don't ensure employability, particularly in cases where a graduate's performance is subpar and internship requirements are not fulfilled. Researchers in the field of higher education face a significant challenge in predicting employability due to the multitude of factors that contribute to this issue. This research presented the methodology for more accurately classifying student data in order to overcome this drawback. Online surveys were used to collect data from Princess Nourahbint Abdulrahman University's College of Computer and Information Sciences information systems (IS) graduates (PNU). The Switching Hierarchical Gaussian Filter (SHGF) is used to preprocess the data at the pre-processing stage. The outcome from the pre-processing is transferred to the feature selection method, which uses Siberian Tiger Optimization (STO) to select the student features. The employed, continuing studies, unemployed, and training are successfully classified using the Multi Fidelity Deep Neural Network. The proposed MFDNN -STO applied to the MATLAB/Simulink platform. To calculate the proposed approach, performance metrics including recall, ROC, computation time, accuracy, precision, sensitivity, and F-score were examined. Higher accuracy of 16.65%, 18.85%, and 17.89%, as well as higher sensitivity of 16.34%, 12.23%, and 18.54%, are achieved by the suggested MFDNN-STO method. The computation time was reduced by 14.89%, 16.89%, and 18.23% as well as 82.37%, 94.47%, and 87.76% in comparison to the existing method.

Keywords: Savitzky Golay (SG) filtering, Siberian Tiger Optimization, using Density Clustering and Graph Neural Network, Employed, Unemployed, Training of Data.

I. INTRODUCTION

Every year, more people complete their higher education and become graduates. Thus, in the Philippines, the employability of graduates continues to be a national concern. According to the DOLE, the country's growing underemployment problem is primarily caused by job mismatches [1]. As of January 2019, the Labor Force Survey in the Philippines revealed a 15.6% underemployment rate and a 5.2% unemployment rate [2]. The Philippines' unemployment rate was 5.7% in 2019, based on the Global Employment Trends study [3] from the International Labour Organisation. Furthermore, according to the PSA data on educational attainment, 20.9% of jobless individuals hold a college degree [4]. The gaps that lead to a mismatch in job skills need to be filled in order to fully realize the Philippines' youthful labor force's potential for productivity. This comprises the number of underemployed and self-employed workers in addition to the number of unemployed college graduates.

Nowadays, most universities consider internships to be mandatory components of their curricula. "An internship is described as "a brief practical work experience in which students receive training and gain experience in a specific field or career area of their interest." [5]. The transition from academia to the workplace is facilitated by internship programs, which serve as a useful means of bridging the knowledge gap between academic requirements and industry demands [6]. A student can apply what they have learned in lectures, combine theory by engaging in an internship programme, individuals can become more acquainted with the workplace, clarify their career goals, and acquire practical experience and employable skills in addition to gaining real-world knowledge through experiential learning [7, 8]. According to research, internships rank among the most important opportunities for experiential learning that increase a graduate's employability [9, 10]. Planning, putting together, and producing a fruitful internship, however, can be difficult [11, 12]. The factors associated with internships that affect students' employability after graduation are not well studied, despite the existence of studies that concentrate on employability prediction [13]. This is partly because there is a dearth of internship data and it is challenging to collect and analyze this data. Very little research has been done to look at how internship programs affect employment prospects; one example of this is [14]. Research predicting a student's employability based on their involvement in internship programs is even less common [15, 16].

^{1*} Lecturer, Academic Affairs Office, Wuhan Business University, Wuhan, HuBei, 430056, China. louisbin_2004@126.com

Copyright © JES 2024 on-line: journal.esrgroups.org

Colleges and universities should work to improve student employment management; they should also give student employment management a bigger role; they should improve student employment guidance so that students know where their future development is headed; and they should support student employment in the future. Thus, in the context of advanced learning, to increase the quality of students, it is critical to focus on the innovation and expansion of employment counselling for university attendees [17, 18]. and, consequently, the caliber of college personnel training [19]. Managing student employment is also important from a practical stand point. There are still issue with college employment data management techniques, despite their positive development in recent decades. In the field of AI, a novel ML technique called profound learning was recently proposed [20]. By training large amounts of data, profound learning can uncover and extract the intricate relationships between the data, increasing classification and prediction accuracy [21]. It's a powerful big data processing technique. Moreover, deep learning models train more quickly than general methods, and as the number of training samples increases, they can demonstrate superior performance growth.

Significant contribution to this work, as outlined here;

- Firstly, information is obtained through an online survey distributed to Information Systems (IS) graduates at PNU's CCIS
- Using a Switching Hierarchical Gaussian Filter (SHGF) eliminate the missing value of Information Systems (IS) department at the CCIS in the PNU employability dataset in the pre-processing segment.
- Using siberian tiger optimization, the pre-processing result is sent into the feature selection procedure to choose the student features.
- The collected features are fed into the Multi Fidelity Deep Neural Network in order to effectively categorize the employed, continuing studies, unemployed and training of data.
- Performance metrics like F-score, recall, ROC, specificity, sensitivity, accuracy, precision, and computation time are examined after the proposed MFDNN-STO technique is implemented.

The rest portions of these manuscripts are coordinated as follows: segment 2 examines a survey of the literature; sector 3 explains the proposed approach; Sector 4 offers the findings and discussions; and sector 5 concludes.

II. LITERATURE REVIEW

Numerous research studies based on the estimation of a student's employability through an internship based on different approaches and factors are available in the literature. A few of them are reviews that were adhered to.

Shi [22] have suggested that an employment management system built on big data and deep learning for students at a college. It draws attention to the inadequacies in college students' employability, which can be strengthened by the contribution of political and ideological education. This essay first provides an overview of the condition of student employment management currently in place and offers solutions to enhance employability in order to assist college students in accurately understanding their employment situation. In order to clarify the connection between the administration of college students' jobs and political and ideological education, the second point of emphasis was placed on the reasons behind the deficiency of this type of education in developing employability. Furthermore, there are particular recommendations about the enhancement of college students' employability. Lastly, the employment satisfaction and rate of employment at colleges and universities were enhanced by the deep learning (DL) recommendation model, which was jointly trained, by establishing a link between the enterprise information and student data correlation.

Saidani et al. [23] have suggested that it predicts a student's employability in the context of an internship using gradient boosting models. The internship course is one of the most crucial ones that enables learners to put their knowledge into practice and get ready for the beginning of a professional career. Nevertheless, internships do not ensure employment, particularly if the graduate does not perform well or did not meet requirements. Because there are so many variables involved in employability prediction, researchers in the field of higher education face a great deal of challenge. Using Gradient Boosting classifiers, context-based employability prediction for students is being introduced in an efficient manner. The employability status prediction processes we have contributed use gradient boosting algorithms to perform context-awareness. Finding the most predictive characteristics influencing graduates' employment prospects is the foundation of student employability prediction.

Yin [24] have suggested that the administration of jobs for Students at a college via deep learning model modeling and big data analysis. This study intends to provide a scientific foundation for managing graduates'

job data by analysing existing employment data, developing a DBN model, and proposing appropriate management techniques. It begins with the large data of employment of college students. According to this study, the DBN model performs better when used to apply the employment data characteristics of college students' management with lower error rates and better accuracy when compared to the BP neural network and linear regression method. Furthermore, the employment rates of college students are primarily determined by the amount of individuals who have completed college and the expansion of the social economy; college employment management platforms will be built using big data technology as a result, providing students with a way to access professional and employment management data.

Veluri et al. [25] have suggested that Deep learning approaches be used in learning analytics to manage educational institutions more effectively. Higher education institutions are becoming more and more like focusing primarily, service providers, on the needs of their students. Enhancing student performance was a university's first priority. Before creating a program to raise the students' performance, it was imperative to evaluate their current circumstances. Administrators in higher education have a difficult time projecting a student's future success. Finding out what influences college students' decisions about their majors was the aim of the research. It will be possible to predict the behavior, mindset, and behavior by utilizing predictive procedures and tools. Raising achievement levels can be accomplished by taking proactive steps to predict student performance in advance. Many attempts to predict student performance have been made in an effort to achieve a high education standard.

Predovic et al. [26] have suggested that a tool for game-based analytics are used to predict behaviours linked to employability skills in order to measure employability after domestic and international internships. The underlying relationships between 33 behavioural descriptors which assess traits like assiduity, social intelligence, learning agility, leadership, and logical reasoning were investigated using exploratory factor analysis. The social component, which focused on how people interact with one another and the outside world, and the cognitive component, which demonstrated the process of learning new information and the motivation to learn, combined to form a two-factor structure. The Cognitive factor were predicted by involvement in an international internship, but not the Social factor. It are the first study that we are aware of that measures employability through international internships using a game-based tool, and it is the first to demonstrate that engaging in foreign internships is linked to cognitive skills as opposed to just social and interpersonal skills.

Bai et al. [27] have suggested that Predictive research for the higher education stage has become a strategic educational asset that has been instrumental in promoting transformation in education. The phrase "educational big data" describes the quickly growing body of educational data that includes psychological states, learning behaviors, and intrinsic characteristics of students. Big data in education has numerous benefits uses in teaching innovation, research management, and educational administration. These programs can be used to suggest jobs, forecast students' academic performance, and help low-income students with financial aid.

Casuat and Festijo [28] have suggested that finding the largely predictive characteristics among undergraduate students' employability signals. Put differently, the objective of the paper is to ascertain, through scientific means, which of the input attributes were most predictive. Via their career centers, Higher Education Institutions (HEI) prioritized finding a scalable and successful strategy for managing student careers. Predictive analytics, a kind of machine learning, aids in accomplishing this objective. It was crucial to be able to foresee which students would be hired and why. The career center and school registrar of the Philippine Institute of Technology furnished the utilized datasets. The datasets comprised 27,000 pieces of data, including the pupils' GPA who were enrolled between 2015 and 2018, the interns' performance evaluation, and the results of their mock job interviews (3,000 observations and 9 attributes). Three approaches were compared in order to identify the most predictive attributes. These included PCA, RFE, and univariate selection (US).

A. Motivation

The general overview of recent studies indicates that, student employability through the internship is an important for the student employability. For scholars working in the field of higher education, predicting employability is becoming increasingly important. Numerous scholars are addressing this issue in the literature regarding various technologies, such as machine learning algorithms,boosting algorithms, Random Forest (RF) algorithm, deep belief network (DBN) RNN, data mining algorithms, Novel Neural Network (NN). Formally, the machine learning prediction processes consist of a collection of activities and their connections. When it comes to applying college students' characteristics to managing employment data, the DBN model offers greater advantages, a lower error rate, and better accuracy. By progressively turning the weak learners from the

previous generation into proficient learners, the boosting algorithm enhances model predictions for learning algorithms. The RF algorithm is used to choose the attributes of businesses and students. Next, it determines if businesses and students are matched. The enterprise resume matching problem is predicted by the RNN algorithms used to process text data. Using data mining algorithms, one can accurately predict an individual's behavior. Organizations are better equipped to allocate personnel and resources in light of this knowledge. Using data from the internet, a novel NN was used to forecast the unemployment rate. The aforementioned technologies have an impact on how difficult and complicated it is to plan, design, and coordinate a successful internship. The employment prediction of college graduates also contains a large amount of error. This study effort was motivated by the shortcomings and issues found in the literature, which primarily address this problem with a lack of approach-based works.

III. PROPOSED METHODOLOGY

In this section, predicting student employability through internship based on big data analysis using a hybrid approach is discussed. The prediction process is highly interested in the contextual data that describes the internship and the student profile. Finding solutions that allow for context-based prediction is therefore essential [29]. This study offers a novel method for estimating the employability status of graduates by using contextual data from the internship and student profile. Proposed methodology is shown in Fig 1. It is divided into four phases: feature selection, classification, pre-processing, and data collection. Consequently, each step's detailed description is provided below.

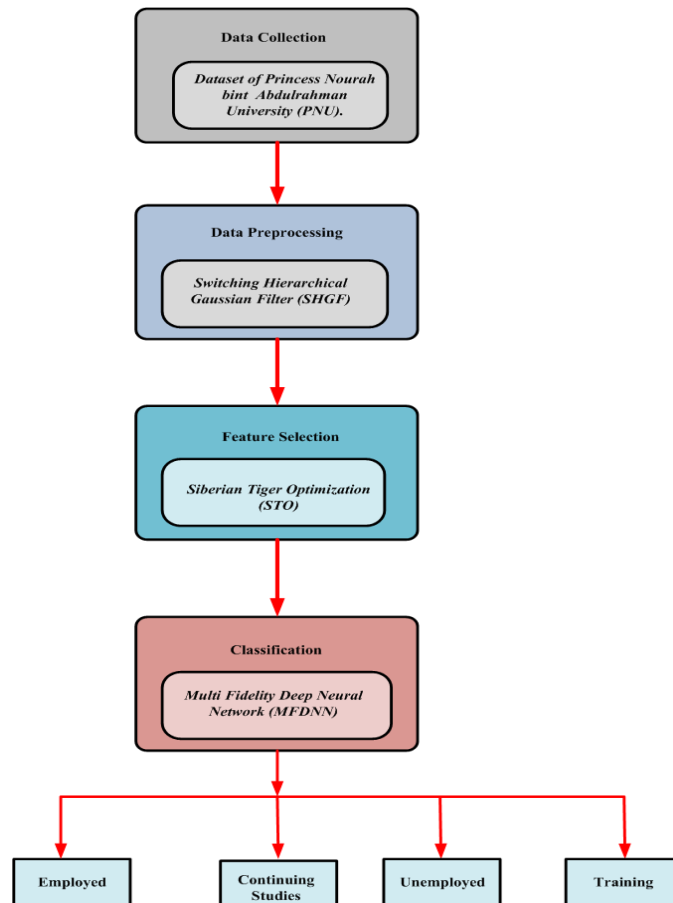


Fig 1: Proposed Approach

A. Data Collection

These three academic years provided the data for this study, from 2019 to 2021. Information Systems (IS) graduates from PNU's CCIS were given an online survey to complete in order to collect data [30].

The survey consisted of three sections: i information about students; ii information about internships; and iii information about jobs. A total of 37 questions were asked. As per the survey's layout, fifty students received an invitation to complete it. After a thorough investigation, no serious problems were discovered, and the students were given an official questionnaire. Two categories of features are identified in this study: features related to

students and features related to internships. GPA, P Certificate, Graduation YS, Cocurricular Activities (Number of Cocurricular Activities), and LinkedIn (Account) are among the features pertaining to students.

The following qualities apply to internships: The factors that affect an internship include its grade, fields, length in training hours or months, kind, number of days per week, technique, rotation and certificate, and contentment with the internship are all represented by the variables IntGrade, IntFields, IntDurationM, and IntDurationH. OrgType (the type of internship organization), OrgSector (the sector of internship organization), OrgJobOffer (the job offer from the internship organization), and OrgRecuitement (recruitment for internship organization) are the features of the organization.

Emp_Status, or the employment status following graduation, is the output feature of the prediction process.

B. Data Pre-Processing using Switching Hierarchical Gaussian Filter (SHGF)

By performing data pre-processing, some respondent records in the employability dataset from the Department of IS at PNU's CCIS have missing values. This section describes the proposed method's detailed pre-processing. Eliminating missing value is the first stage in the pre-processing process. To apply the SHGF for data analysis, once a suitable dataset is obtained, the first step is data cleaning and preparation. To maintain the quality of the dataset, data cleaning entails locating and addressing outliers, missing values, and inconsistencies [31]. The data can be prepared for further analysis by using a variety of techniques, including imputation, filtering, and normalization. After data cleaning, the next step is to apply the SHGF. Applying SHGF to the cleaned data involves setting up the model parameters, selecting appropriate state and parameter transition models and using Bayesian estimation techniques for interface. Once the SHGF analysis is performed, researchers can gain insights into the underlying patterns, trends or states within the data. The results obtained from the SHGF analysis can then be used for various purposes, like prediction, decision-making or further research. As always, when conducting any analysis, it is crucial to validate the results, interpret them correctly, and consider the limitations and assumptions of the chosen. Additionally, researchers should keep abreast of the latest developments in SHGF or any other data analysis techniques beyond my last update to ensure they are using the most up-to-date methodologies for their research.

The state transition model represents how the underlying state evolves over time. In the context of SHGF, this could be a linear or nonlinear model. For a simple linear model, it can be represented as follows in equation (1)

$$x_{\{t\}} = a * x_{\{t-1\}} + w \tag{1}$$

In this case, the state vector at t-1 is indicated by $x_{\{t-1\}}$, the state vector at time (t) is indicated by $x_{\{t\}}$. A is represent as the state transition matrix that models how the state evolves, w is the process noise or noise, assumed to be Gaussian with zero mean. The observation model relates the observed data to the underlying state. In SHGF, this can be a linear or nonlinear model, depending on the application. For a simple linear model, it can be represented in equation (2)

$$y_{\{t\}} = H * x_{\{t\}} + v \tag{2}$$

where $y_{\{t\}}$ the observed data or measurement at time t , H is indicated as the observation matrix that maps the state space to the observation space, v is the observation noise, thought to have a zero mean and be Gaussian. SHGF utilises Gaussian distributions to model the uncertainties in the state transition and observation processes. These distributions are commonly represented in equation (3)

$$p(x_{\{t\}} \setminus x_{\{t-1\}}) \sim N(A * x_{\{t-1\}}, Q) \tag{3}$$

Here $p(x_{\{t\}} \setminus x_{\{t-1\}})$ is denoted as the state's probability distribution at a time, t given the state at time $t - 1$, Q is represent as the covariance matrix representing the process noise and is determined in equation (4)

$$p(y_{\{t\}} \setminus x_{\{t\}}) \sim N(A * x_{\{t\}}, R) \tag{4}$$

where $p(y_{\{t\}} \setminus x_{\{t\}})$ | is indicated as given the state at time t , the observation's probability distribution at time t , R is the covariance matrix representing the observation noise. SHGF employs Bayesian estimation to update the beliefs about the underlying state based on new observations. The estimation process involves computing the posterior distribution of the state given the observations using Bayes' theorem as shown in equation (5)

$$p(x_{\{t\}} \setminus y_{\{1:t\}}) = (p(y_{\{t\}} \setminus x_{\{t\}}) * p(x_{\{t\}} \setminus y_{\{1:t-1\}})) / p(y_{\{t\}} \setminus y_{\{1:t-1\}}) \tag{5}$$

here $p(x_{-t} | y_{-1:t})$ is indicated as the posterior distribution of the state at time t given all the observations up to time t , $p(y_{-t} | y_{-1:t-1})$ is the likelihood term that acts as a normalizing constant.

Finally, before applying the Switching Hierarchical Gaussian Filter (SHGF) for further analysis, thorough cleaning and preparation of the dataset are essential to enhancing data quality and suitability. Proper data cleaning can help address missing values, outliers, and inconsistencies, ensuring that the SHGF algorithm can work effectively and provide meaningful insights.

The data is preprocessed and the employability dataset contains missing values are rejected and the preprocessed data is fed into the feature selection.

C. Feature Selection using Siberian Tiger Optimization

More distinguishing features are found and superfluous and irrelevant features are eliminated thanks to this process. We refer to a feature as irrelevant if it contains no information about the different classes. It is considered a redundant feature if it exhibits a strong correlation with other features and reduces accuracy.

1) Siberian Tiger Optimization

Siberian tiger's natural behaviours are intelligent processes that can inspire the design of new metaheuristic algorithms [32]. Brown and black bears engage in combat with Siberian tigers over prey and in self-defense. In this encounter, the bear is ambushed by the Siberian tiger, which then strikes it from above, with one forepaw, stop it; with the other, grab its throat and bite its spine to kill it. STO possesses outperformed competing algorithms in handling optimization applications and has demonstrated superior power in exploring, exploitation, and balancing these tasks.

Step 1: Initialization

Initialize the input parameters like magnetic field strength, nanoparticle concentration, etc.

$$y_{i,j} = lb_j + r_{i,j} \cdot (ub_j - lb_j), \quad i = 1, 2, \dots, D$$

$$j = 1, 2, \dots, m \tag{6}$$

here ub_j, lb_j are the upper and lower bound of the j th problem variable, $r_{i,j}$ represent as the arbitrary value in the interval $[0, 1]$, $y_{i,j}$ specifies j th dimension of y_i in the search space, m indicates the count of problem variable.

Step 2: Random Generation

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_N \end{bmatrix}_{D \times m} = \begin{bmatrix} y_{1,1} & \cdots & y_{1,j} & \cdots & y_{1,m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{i,1} & \cdots & y_{i,j} & \cdots & y_{i,m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{N,1} & \cdots & y_{N,j} & \cdots & y_{N,m} \end{bmatrix}_{D \times m} \tag{7}$$

where Y is denoted as the location matrix of the Siberian tiger population, Y_i is indicated as the j th Siberian tiger, and D is represent as the total count of Siberian tigers.

Step 3: Fitness Calculation

The primary objective that determines the fitness is explained by,

$$Fitnessfunction = \text{optimizewithFeatureSelection} \tag{8}$$

Step 4: Exploration Phase: Pray Hunting

Siberian tigers are apex attackersthat hunt and eat a range of targets.As a result, in the early period, STO members are updated using a imitation of the pursuingtactics used by Siberian tigers. In these strategy, the Siberian tiger chooses its victim, attacks it, and then chases after the animal to kill it. Therefore, two steps are used to imitate the prey hunting phase.

At first stage, the population members' positions are reorganized based on their selection and attack of prey. The positions of the STO members alter drastically and abruptly during this period. In equation (9), the set of prey places is displayed.

$$NN_i = \{Y_k | k \in \{1, 2, \dots, D\} \wedge F_k < F_i\} \cup \{Y_{best}\}, \tag{9}$$

here Y_{best} is denoted as the best candidate elucidation and The total number of members of the STO is represented by D . Using equation (10), a simulation of the victim's attack is used to plan the new position of one member (TN_i) from this set of NN_i , which is randomly chosen to be the target of the i th seberian tigers.

$$y_{i,j}^{N1S1} = y_{i,j} + r_{i,j} \cdot (TN_{i,j} - I_{i,j} \cdot y_{i,j}), \quad (10)$$

$$i = 1, 2, \dots, D \text{ and } j = 1, 2, \dots, m$$

here $TN_{i,j}$ indicated the j th dimension of TN_i , $y_{i,j}^{N1S1}$ indicated the j th dimension of y_i^{N1S1} . If a STO member's newly calculated position increases the goal function value, it is acceptable. equation (11) is used to model this process.

$$Y_i = \begin{cases} Y_i^{N1S1}, & F_i^{N1S1} < F_i; \\ Y_i, & \text{else,} \end{cases} \quad (11)$$

where Y_i^{N1S1} , the i th member's noval position, as determined by the initial phase of STO's first stage, is designated by the symbol F_i^{N1S1} , which represents the i th member's objective function value.

At the second step, the population members' positions are updated in accordance with the chase method. In initial, equation (12) determines the siberian tigers' new location in relation to the assault site to replicate the process of pursuing something. Then according to equation (13),

$$y_{i,j}^{N1S2} = y_{i,j} + \frac{r_{i,j} \cdot (ub_j - lb_j)}{t}, i = 1, 2, \dots, D, \quad (12)$$

$$j = 1, 2, \dots, m, \text{ and } t = 2, \dots, T,$$

$$Y_i = \begin{cases} Y_i^{N1S2}, & F_i^{N1S2} < F_i; \\ Y_i, & \text{else,} \end{cases} \quad (13)$$

where F_i^{N1S2} indicates its value of objective function, and t is denoted as the iteration counter of the algorithm, $y_{i,j}^{N1S2}$ is its j th dimension, Y_i^{N1S2} is denoted as the noval location of the i th Siberian tigers according to the Phase two of the STO's phase two.

Step 5: Exploitation Phase: Fighting with a Bear

Siberian tigers demonstrate how this species engages in combat with brown and black bears in order to protect its target and ensure its survival. As a result, in the 2nd stage STO members are modernized using a simulation of the Siberian tigers' battle strategy against bears. The Siberian tiger ambushes the bear in this encounter before hitting. The conflict between the Siberian tigers and the bear is portrayed in 2 stages. The remainder of the population are initially thought of as a set of bears in order to simulate the harass of the i th Siberian tiger. First, using equation (14), the new position of the i th STO member is determined.

$$y_{i,j}^{N2S1} = \begin{cases} y_{i,j} + r_{i,j} \cdot (y_{k,j} - I_{i,j} \cdot y_{i,j}), & F_k < F_i; \\ y_{i,j} + r_{i,j} \cdot (y_{k,j} - I_{i,j} \cdot y_{i,j}), & \text{else,} \end{cases} \quad (14)$$

where y_i^{N2S1} is the new member's position based on the 1st phase of the 2nd stage of STO, $y_{i,j}^{N2S1}$ is its j th dimension, $y_{k,j}$ is denoted as the j th dimension of a location of bear, $j = 1, 2, \dots, m$, k is denoted as arbitrarily certain from the set $\{1, 2, \dots, i-1, i+1, \dots, D\}$, F_k is represent as the significance of the bear's objective function. Equation (15) states that the previously determined values of the respective members are replaced with the newly calculated values.

$$Y_i = \begin{cases} Y_i^{N2S1}, & F_i^{N2S1} < F_i; \\ Y_i, & \text{else,} \end{cases} \quad (15)$$

where F_i^{N2S1} is the objective function value of Y_i^{N2S1} .

At the second step, the population members' positions are changed in accordance with a simulation of conflict that occur during warfare. First, a arbitrary point close to the fight's location is determined using

equation (16) in order to replicate this behaviour. If the charge of the objective function is better as per equation (17), this new location is then suitable for the update procedure.

$$y_{i,j}^{N2S2} = y_{i,j} + \frac{r_{i,j}}{t} (ub_j - lb_j), i = 1, 2, \dots, D, \tag{16}$$

$$j = 1, 2, \dots, m, \text{ and } t = 2, \dots, T,$$

$$Y_i = \begin{cases} Y_i^{N2S2}, & F_i^{N2S2} < F_i; \\ Y_i, & \text{else,} \end{cases} \tag{17}$$

where, F_i^{N1S2} indicates its value of objective function, and t is denoted as the algorithm's iteration counter, $y_{i,j}^{N2S2}$ is its j th dimension, y_i^{N2S2} is represented as the new Siberian tiger location determined by the 2ndstage of the 2ndphase of STO.

Step 6: Update the Best solution

The iteration of a STO ends when all search areas for Siberian tigers have been updated and established based on the first and second stages.

Step 7: Termination

If the best solution is discovered after reviewing the stopping criteria, the process ends; if not, go on to step 3.

Professional certificates, graduation year, semester, and linked In account are the selected features of the given dataset. The feature selection is transferred to the classifier; the MFDNN is used for classifying the selected data. The classify method is shown below;

D. Classification by Multi Fidelity Deep Neural Network (MFDNN)

The MFDNN blends different fidelity information to construct accurate responses. The primary difficulty in multi-fidelity modelling is utilizing the relationship between information with different levels of fidelity [33]. The auto-regressive approach is used and expressed as follows in the current work:

$$z_H = \rho(x)z_L + \delta(x), \tag{18}$$

where, z_L and z_H indicate the fidelity the data, from low to high, correspondingly; $\rho(x)$ estimates the correlation between two variables using a scalar variable $\{ z_H, z_L \}$, and δ is the corresponding noise. The high and low fidelity data's linear correlation may be effectively managed with the help of this method. It is crucial to understand that these two fidelity metrics have a nonlinear relationship.

$$z_H = \alpha f_l(y_H, z_L) + (1 - \alpha) f_{nl}(y_H, z_L), \alpha \in [0, 1], \tag{19}$$

In this case, f represents a mapping between low-fidelity neural network forecasting z_L and high-fidelity neural network forecasting z_H using high-fidelity input y_H . f is divided into 2 parts: f_{nl} represents nonlinear terms, and f_l represents the linear term. Building a correlation between the provided data and the higher and lower fidelity data is a new hyper-parameter α . Greater linear correlation among low/ high -fidelity data is indicated by a larger value of α .

The first neural network NN1 (y_L, θ) serves as an estimate for the low-fidelity data. The 2nd and 3rd neural networks, NNH1 ($y, y_L, \beta_i, i = 1, 2$) are employed to roughly estimate the correlation among low/ high -fidelity data, where NNH1 and NNH2 correspond to the linear and nonlinear correlations, respectively $\theta, \alpha, \beta_i, (i = 1, 2)$ are the MFDNN model's undiscovered hyper-parameters, where θ and $\beta_i, (i = 1, 2)$ represent the biases and weights of neural networks.

Finally, the MFDNN processing that divides the data into categories such as training, jobless, continuing education, and employed

IV. RESULT AND DISCUSSION

The experimental result of the predicting student employability through internship based on data analysis technique using MFDNN-STO method is discussed in this section. MATLAB/Simulink is used to perform the simulations. MATLAB is used to simulate the proposed method using a number of performance metrics. The results of MFDNN analyzed using RNN, DBN, and LGBN, three of the existing techniques.

A. Presentation metrics

Included is a comparison of the performance metrics, including sensitivity, specificity, ROC, f1-score, and computational time. To evaluate the performance measurements, the confusion matrix is required. To scale the confusion matrix, one must comprehend the true negative (tn), false negative (fn), true positive (tp) values and false positive (fp)

1) Accuracy

The total count of occurrences within the dataset is how it is defined. The outcome is a matrix that describes how well the model performs in each class. Consequently, equation establishes it eqn (20),

$$Accuracy(acc) = \frac{tp + tn}{tp + tn + fp + fn} \tag{20}$$

2) Sensitivity

Sensitivity is the name for a true positive rate or recall. Equation (21), which calculates the sensitivity,

$$Sensitivity(sen) = \frac{tp}{tp + fn} \tag{21}$$

3) Precision

Equation (22) is used to calculate the precision, it's referred to as real positive predictive principles.

$$Precision = \frac{tp}{tp + fp} \tag{22}$$

4) F-score

This produces the value of positive prediction in addition to the sensitivity's harmonic mean. Baye's theorem states that the positive predictive value (pv).

$$pv = \frac{sen \times p}{(sen \times p + (1 - spe) \times (1 - p))} \tag{23}$$

$$p = \frac{tp + fn}{tp + fp + fn + tn} \tag{24}$$

Equation (25), therefore, is used to calculate the F-score.

$$F - Scorevalue = 2 \times \frac{sen \times pv}{sen + pv} \tag{25}$$

5) Recall value

Equation (26) represents the value of recall,

$$Recall = \frac{pt}{(tp + fn)} \tag{26}$$

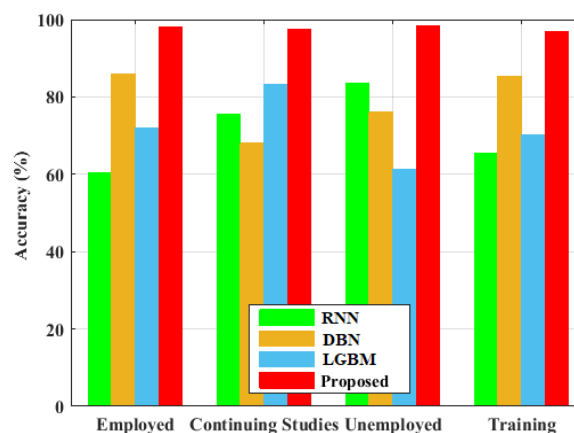


Fig 2: The comparison of accuracy values between the proposed and existing systems.

The comparison of accuracy values between the proposed and existing systems is displayed in Fig 2. The performance of the proposed method outcome in accuracy is 50.56%, 20.76%, 35.97% advanced for the

classification of employed, 20.46%, 35.58%, 23.54% advanced for the classification of continuing studies and 21.45%, 30.76%, 18.43% higher for the classification of unemployed, 21.44%, 30.86%, 15.43% higher for the classification of training when evaluated to the existing RNN, DBN, and LGBM models correspondingly.

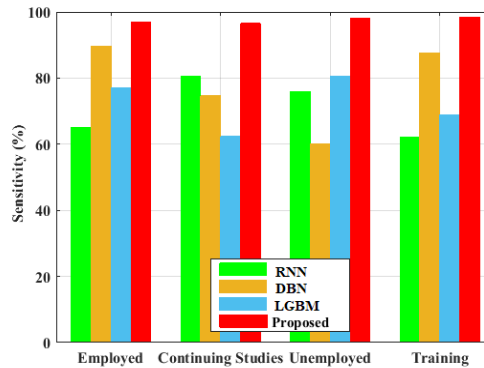


Fig 3: A comparison of the proposed and existing systems' sensitivity values.

A comparison of the proposed and existing systems' sensitivity values is displayed in Fig 3. The performance of the proposed technique results in precision that is 30.56%, 21.76%, 35.97%, higher for the classification of employed, 21.46%, 33.58%, 23.54% higher for the classification of continuing studies and 21.45%, 30.76%, 18.43% higher for the classification of unemployed, 20.46%, 35.58%, 23.54% higher for the classification of training when evaluated with the existing RNN, DBN, and LGBM models correspondingly.

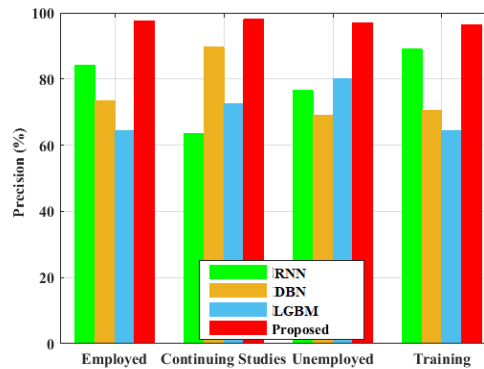


Fig 4: The precision value comparison between the proposed and existing systems.

The precision value comparison between the proposed and existing systems is illustrated in Fig 4. Here, a direct comparison with proposed approaches is offered to show how the suggested method's precisions are higher. The proposed technique provides for more extensive analyses of a proposed and has higher precision than existing methods due to its wider consideration of factors. The performance of the proposed technique results in precision that are 30.56%, 21.76%, 35.97%, higher for the classification of employed, 21.46%, 33.58%, 23.54% higher for the classification of continuing studies and 21.45%, 30.76%, 18.43% higher for the classification of unemployed, 20.44%, 30.86%, 15.43% higher for the classification of training when evaluated to the existing RNN, DBN, and LGBM models correspondingly.

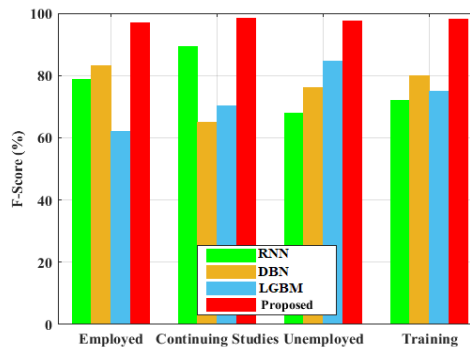


Fig 5: The F-score value comparison between the proposed and existing systems.

The F-score value comparison between the proposed and existing systems is illustrated in Fig 5. The performance of the proposed technique results in precisions that are 22.56%, 21.76%, 33.97%, higher for the classification of employed, 21.46%, 33.58%, 23.54% higher for the classification of continuing studies and 21.45%, 30.76%, 18.43% higher for the classification of unemployed, 20.44%, 30.86%, 15.43% higher for the classification of training when evaluated with the existing RNN, DBN, and LGBM models correspondingly.

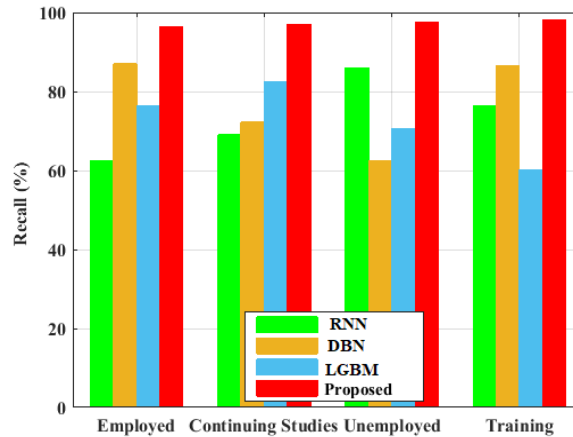


Fig 6: The comparison of the proposed and existing systems' recall values.

The comparison of the proposed and existing systems' recall values is displayed in Fig 6. The performance of the proposed technique results in recall that are 23.56%, 22.76%, 31.97% higher for the classification of employed, 22.46%, 31.58%, 22.54% higher for the classification of continuing studies and 22.45%, 29.76%, 17.43% higher for the classification of unemployed, 19.44%, 29.86%, 14.43% higher for the classification of training when evaluated to the existing RNN, DBN, and LGBM models correspondingly.

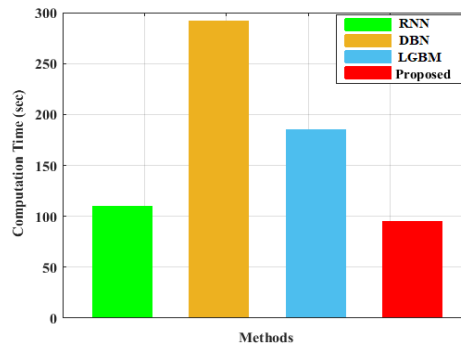


Fig 7: Analyses of computation time using proposed and existing methods.

Analyses of computation time using proposed and existing methods is displayed in Fig 7. The proposed MFDNN-STO is evaluated against the RNN, DACNN, and PCSANN existing techniques. In comparison to existing methods such as RNN, DACNN, and PCSANN, the proposed DCGNN-STO method yields 5.34%, 10.11%, and 10.26% lower computational time, correspondingly. when compared to the RNN, DBN, and LGBM models that are currently in use, correspondingly.

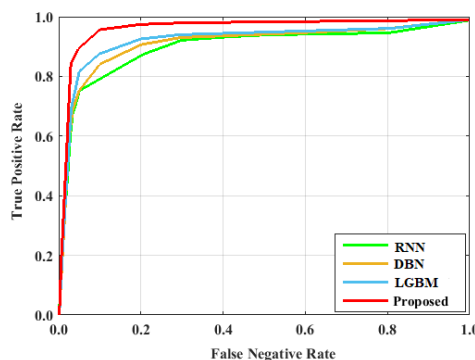


Fig 8: True positive rate comparison between the proposed and existing methods.

True positive rate comparison between the proposed and existing methods is displayed in Fig 8. A ROC curve is produced in a two-dimensional space by illustrating the false positive rate on the x-axis and the rate of true positives on the y-axis. The system's great ability to diagnose breast masses, which can be used to lower the death rate, is demonstrated by a low false positive rate. The ROC of the proposed MFDNN-STO technique is assessed against the RNN, DBN, and LGBN existing methods. Subsequently, the proposed MFDNN-STO yields an AUC greater than the existing methods in 3.49%, 6.45%, and 6.78%, which include RNN, DBN, and LGBN models correspondingly.

V. CONCLUSION

An essential first step in data is collected via an online survey given to IS department graduates at PNU's CCIS. The employability data are pre-processed using the Switching Hierarchical Gaussian Filter (SHGF). Following pre-processing, the feature selection process uses Siberian Tiger Optimization (STO) to choose employability data. The selected features are then provided to the MFDNN to efficiently classify the employed, continuing studies, unemployed and training of student. The MATLAB platform is used to evaluate the proposed strategy and compare it to other current strategies. The proposed approach is investigated in a variety of scenarios, including recall, ROC, computation time, F-score, p accuracy, precision, and sensitivity. The final step of the system uses the MFDNN classifier, which has increased its accuracy to 98% due to its precise identification of the recommended optimum region expanding technique. This shows that the classifier is capable of accurately identifying employed, continuing studies, unemployed, and student training.

REFERENCE

- [1] Othman, Z., Shan, S. W., Yusoff, I., & Kee, C. P. (2018). Classification techniques for predicting graduate employability. *International Journal on Advanced Science, Engineering and Information Technology*, 8(4-2), 1712-1720.
- [2] Crasta, L. C., & Shailashri, V. T. (2023). A Systematic Review on the Employability Prediction Model for the Management Students. *International Journal of Case Studies in Business, IT and Education (IJCSBE)*, 7(1), 1-25.
- [3] Khadragy, S. (2020). *Data Analytics: Adaptive Network-based Fuzzy Inference System for prediction of computer science graduates' employability* (Doctoral dissertation, The British University in Dubai).
- [4] Xu, L., Zhang, J., Ding, Y., Sun, G., Zhang, W., Philbin, S. P., & Guo, B. H. (2022). Assessing the impact of digital education and the role of the big data analytics course to enhance the skills and employability of engineering students. *Frontiers in Psychology*, 13, 974574.
- [5] Xiaoying, W. (2020). The construction of employment and entrepreneurship service system for university students based on big data. *Asia-pacific Journal of Convergent Research Interchange*, 10, 203-213.
- [6] Tentama, F., & Yusantri, S. (2020). The Role of Entrepreneurial Intention in Predicting Vocational High School Students' Employability. *International Journal of Evaluation and Research in Education*, 9(3), 558-563.
- [7] Jha, S., Jha, M., & O'Brien, L. (2018, December). A step towards big data architecture for higher education analytics. In *2018 5th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)* (pp. 178-183). IEEE.
- [8] Pinto, L. H., & Pereira, P. C. (2019). 'I wish to do an internship (abroad)': investigating the perceived employability of domestic and international business internships. *Higher Education*, 78, 443-461.
- [9] Jung, J., & Lee, S. J. (2017). Impact of internship on job performance among university graduates in South Korea. *International Journal of Chinese Education*, 5(2), 250-284.
- [10] Bossler, M., & Wegmann, J. (2019). The German generation internship and the minimum wage introduction: evidence from big data. *Applied Economics*, 51(16), 1730-1747.
- [11] Johnson, M., Jain, R., Brennan-Tonetta, P., Swartz, E., Silver, D., Paolini, J., ... & Hill, C. (2021). Impact of big data and artificial intelligence on industry: developing a workforce roadmap for a data driven economy. *Global Journal of Flexible Systems Management*, 22(3), 197-217.
- [12] Brown, P., & Souto-Otero, M. (2020). The end of the credential society? An analysis of the relationship between education and the labour market using big data. *Journal of Education Policy*, 35(1), 95-118.
- [13] Tentama, F., & Nur, M. Z. (2021). The Correlation between Self-Efficacy and Peer Interaction towards Students' Employability in Vocational High School. *International Journal of Evaluation and Research in Education*, 10(1), 8-15.
- [14] Pan, J., Guan, Y., Wu, J., Han, L., Zhu, F., Fu, X., & Yu, J. (2018). The interplay of proactive personality and internship quality in Chinese university graduates' job search success: The role of career adaptability. *Journal of Vocational Behavior*, 109, 14-26.
- [15] Cahan, E. M., Hernandez-Boussard, T., Thadaney-Israni, S., & Rubin, D. L. (2019). Putting the data before the algorithm in big data addressing personalized healthcare. *NPJ Digital Medicine*, 2(1), 78.

- [16] Ebner, K., Soucek, R., & Selenko, E. (2021). Perceived quality of internships and employability perceptions: the mediating role of career-entry worries. *Education+ Training*, 63(4), 579-596.
- [17] Viji, D., Dhanka, S., Binda, M. B., & Thomas, M. (2024). Hybrid STO-IWGAN method based energy optimization in fuel cell electric vehicles. *Energy Conversion and Management*, 305, 118249.
- [18] Singh, N. M., & Sharma, S. K. (2023). An efficient automated multi-modal cyberbullying detection using decision fusion classifier on social media platforms. *Multimedia Tools and Applications*, 1-29.
- [19] Persaud, A. (2021). Key competencies for big data analytics professions: A multimethod study. *Information Technology & People*, 34(1), 178-203.
- [20] Huang, X., Cao, J., Zhao, G., & Long, Z. (2022). The employability and career development of finance and trade college graduates. *Frontiers in psychology*, 12, 719336.
- [21] He, S., & Li, Y. (2023). Design and application of college student management system based on big data technology. *Wireless Communications and Mobile Computing*, 2023.
- [22] Shi, Q. (2023). The Employment Management for College Students based on Deep Learning and Big Data. *IEEE Access*.
- [23] Saidani, O., Menzli, L. J., Ksibi, A., Alturki, N., & Alluhaidan, A. S. (2022). Predicting student employability through the internship context using gradient boosting models. *IEEE Access*, 10, 46472-46489.
- [24] Yin, Y. (2023). Big Data Analysis and Modeling Method of College Student Employment Management Based on Deep Learning Model. *International Journal of Web-Based Learning and Teaching Technologies (IJWLTT)*, 18(2), 1-15.
- [25] Veluri, R. K., Patra, I., Naved, M., Prasad, V. V., Arcinas, M. M., Beram, S. M., & Raghuvanshi, A. (2022). Learning analytics using deep learning techniques for efficiently managing educational institutes. *Materials Today: Proceedings*, 51, 2317-2320.
- [26] Predovic, D., Dennis, J., & Jones, E. (2022). International internships and employability: a game-based assessment approach. *Higher Education Research & Development*, 41(4), 1231-1246.
- [27] Bai, X., Zhang, F., Li, J., Guo, T., Aziz, A., Jin, A., & Xia, F. (2021). Educational big data: Predictions, applications and challenges. *Big Data Research*, 26, 100270.
- [28] Casuat, C. D., & Festijo, E. D. (2020, February). Identifying the most predictive attributes among employability signals of undergraduate students. In *2020 16th IEEE international colloquium on signal processing & its applications (CSPA)* (pp. 203-206). IEEE.
- [29] Stanton, W. W., & Stanton, A. D. A. (2020). Helping business students acquire the skills needed for a career in analytics: A comprehensive industry assessment of entry-level requirements. *Decision Sciences Journal of Innovative Education*, 18(1), 138-165.
- [30] Pattanapairoj, S., Nitisiri, K., & Sethanan, K. (2021). A gap study between employers' expectations in thailand and current competence of master's degree students in industrial engineering under industry 4.0. *Production Engineering Archives*, 27(1), 50-57.
- [31] Wu, J. M. T., Tsai, M. H., Huang, Y. Z., Islam, S. H., Hassan, M. M., Alelaiwi, A., & Fortino, G. (2019). Applying an ensemble convolutional neural network with Savitzky–Golay filter to construct a phonocardiogram prediction model. *Applied Soft Computing*, 78, 29-40.
- [32] Trojovský, P., Dehghani, M., & Hanuš, P. (2022). Siberian tiger optimization: A new bio-inspired metaheuristic algorithm for solving engineering optimization problems. *Ieee Access*, 10, 132396-132431.
- [33] Xiong, S., Li, B., & Zhu, S. (2023). DCGNN: A single-stage 3D object detection network based on density clustering and graph neural network. *Complex & Intelligent Systems*, 9(3), 3399-3408.