Missing Data Prediction using Correlation Genetic Algorithm and SVM Approach

Aysh Alhroob[1], Wael Alzyadat[2], Ikhlas Almukahel[3], Hassan Altarawneh[4]
Department of Software Engineering, Faculty of Information Technology Isra University, Jordan[1,3]
Department of Software Engineering Faculty of Science and Information Technology Al-Zaytoonah University, Jordan[2]
Department of Computer Science Faculty of Information Technology Middle East University, Jordan[4]

***Abstract*—Data exists in large volume in the modern world, it becomes very useful when decoded correctly to inform decision making towards tackling real word issues. However, when the data is conflicting, it becomes a daunting task to get obtain information. Working on missing data has become a very important task in big data analysis. This paper considers the handling of the missing data using the Support Vector Machine (SVM) based on a technique called Correlation-Genetic Algorithm-SVM. This data is to be subjected to the SVM classification technique after identifying the attribute's correlation and application of the genetic algorithm. The application of the correlation enables a clear view of the attributes which are highly correlated within a particular dataset. The results indicate that apart from the SVM, the application of the proposed hybrid algorithm produces better outcomes identification rate and accuracy is considered. The proposed approach is also compared with depicts the Mean Identification rate of applying the neural network, the result indicate a consistent accuracy hence making it better.**

***Keywords*—Missing data; Support Vector Machine (SVM); genetic algorithm; hybrid algorithm; correlation**

I. Introduction

Data missing is the most common issue in various real worlds because it affects taking a timely decision using the acquired data. This research addresses missing data issues of data preprocessing that can have a significant impact on generalization performance of classification accuracy towards meaningful data. Various dataset suffer from an unavoidable problem of missing values for many reasons such as not enough data in report results; missing in industrial experiment, or failures automatic machine while collecting data [1] medicinal dataset contains missing data because some patient's record needs some critical value, not all possible tests to investigate it [2].

Data missing may happen at two stages; during training time as training data or at the prediction time while testing the data. The machine learning algorithms are mainly concerned with the identification of the missing values at the training time with less focus on missing values during prediction time. There are various techniques for treating missing data, examples include imputation techniques, ignoring techniques, and model-based techniques. The ignoring technique includes complete case analysis, which involves analyzing the case to have any missing data in any of the variables. The particular case is omitted from the analysis part. Another technique in ignoring is pair-wise deletion in which each of the features is considered and the value missing in any field is not much minded. Treating missing data requires thorough analysis process involving estimation of missing value without losing the statistical perspective of the dataset. These two criteria are contradictory and use the information from the partially completed data and at the same time maintaining the statistical perspective of the dataset while imputing the missing values [3]. Some techniques have been discussed to handle of missing data[3], such as remove cell containing missing data other using imputation with appropriate values. The main difference between the two approaches is that, removing missing data is more suitable with small number of instances to avoid information decrease while Imputation methods can be practical with big data and large missing value. Consequently, imputation methods are a accepted approach dealing with missing value.

Correlation is a technique which identifies the relationship between variables. The correlation factor helps in identifying the suitable relation between the variables of a particular dataset. The support vector machine (SVM) is a supervised learning technique which initially helped in two-class classification problem. The kernel functions may also be applied to optimize the parameters. Given a set of training data, SVM produces optimum hyper plane by using the concept of supervised learning. Basically a hyper plane is one which acts as a line that plays an important role in dividing the plane into two parts which belongs to each of the class. The SVM plays a drastic role when there is a clear-cut division of the two classes along X and Y plane. When there is no clear discrimination of the two classes through a particular line, then there is a need to use the third axis Z. There arises the use of kernals. Therefore by using some tuning factors in support vector machine and by changing them according to the problem enables to achieve non-linear classification. This type of classification helps in achieving higher accuracy in limited amount of time. Kernal plays a very important role in learning the hyper plane in SVM which helps in changing the problem using linear algebra. The SVM plays a major role in text categorization removing the need for labeled training data. Image classification is also possible through SVM which provides higher accuracy rate than existing systems. Image segmentation also has the usage of SVM in it.The SVM helps in classification of proteins in biological science and also enables recognition of the handwritten text. SVM also has few disadvantages. The SVM algorithm avoids probability estimation on data which are stable. The input data needs to be fully labeled. The applicability of SVM is more towards two-class problem and further multi-class structure needs to be looked upon.

The genetic algorithm has the basic steps of selection of

population, crossover and mutation. The fitness function determines the quality of the individual. Fitness passed individuals are inherited to another generation. The genetic algorithm initially originates with a set of solutions and later variants them for different generations. For increasing the performance of the algorithm, random search is performed on the old data for new search items. Therefore genetic algorithm allows global search thereby trying to improve the global optimum through various available solutions.

The genetic algorithm has the necessary steps for selection of population, crossover, and mutation. The fitness function determines the quality of the individual, and individuals who pass fitness test are inherited to another generation. The genetic algorithm initially originates with a set of solutions and later variants them for different generations. For increasing the performance of the algorithm, a random search is performed on the old data for new search items. The genetic algorithm creates an opportunity for global search hence improves global optimum through the available solutions. The genetic algorithm tries to identify the attributes with the missing values [4]. Once the attribute has been identified, it engages in finding domain values of missing data values. Values of the missing attributes are then replaced with identified domain values such that possible set of domain values are identified for the missing attribute values. A similar concept applies to all attributes with missing values. With an overall bunch of arrived values, the set of values are chosen. Crossover on the set of selected instances is made. The fitness function is determined and validation is done against it. This helps in the determination of classification accuracy on the decision tree [5] [6] . If the selected instance is classified, then the substituted values are classified or else they are deleted. The process is repeated until a bunch of values is obtained. The proposed paper tries to address the missing data using the concept of correlation, which identifies the relation. Then the genetic algorithm and SVM are applied to handle the missing data and efficiently classify the data.

The paper is organized as follows: Section II deals with the related works in handling missing data, Section III is the proposed work and Section IV deals with the implementation and Section V deals with results and discussion and section VI deals with conclusion.

II. Related Work

Handling missing data is very important in term of use these data. Many Techniques used to optimize the data findings and use. Optimization and Machine Learning algorithms are used to enhance the data processing. The Genetic Algorithm (GA) [7] was used to optimize the initial weight and threshold values of support vector machine. The proposed GA-SVM was used to forecast the CO2 emissions of Beijing [8]. The factors contributing to this was identified to be residential growth, economic factors and the CO2 emissions were found to be more than 0.5. The cancer data is classified using support vector machine and genetic algorithm[9] to find the better accuracy in classification. Radial basis and polynomial kernel [10] function are used in this proposed technique. The proposed technique is compared to the existing techniques based on the runtime also.

In model selection using support vector machine [11], genetic algorithm is being used. The fitness function is calculated and various kernel parameters [12] [13] are determined. The proposed model selection technique is applied on four datasets to observe if it satisfies the criteria. The proposed estimator outperforms giving best fitness criterion that yielded more models. Authors in paper [14] proposed a genetic algorithm for optimizing the parameters of support vector machine. This involves image classification based on object-oriented classification. The proposed system is compared with the grid algorithm and found to be superior in terms of time and accuracy factors.

For the purpose of identifying the damage on the bridge, support vector machine along with genetic algorithm which is customized to get best kernel parameters. The proposed GA-SVM [15] is compared with other back propagation techniques to arrive at the best technique. With the error rates of other technique, it is being concluded that the proposed technique has higher accuracy rate in finding the damage. The least square SVM [16] technique is being proposed which helps in making the complex problem to linear regression one. Then by applying genetic algorithm over this LS-SVM [17], optimal parameters [18] [19] are obtained. The proposed system is compared with other existing systems like artificial neural network and it is found that the LS_SVM based system perform far better than that.

The classifier works in [23], presents how a classifier works if there are missing values in the data. Initially non-parametric technique is used for the data processing. But it narrows to simpler SVM if no missing data is present in the data. Further an analysis of Least square SVM [4] [24] is done to understand the classifier better.

The work in paper [1], is based on the objective to identify the missing rate in a selective manner. The proposed technique helps in achieving a good Mean Identification Rate (MIR) through less imputation method. By understanding the technique, the proposed method is evaluated for the parameter to check if the system is working properly. The paper [2]is based on the functional dependency related technique which is targeted with machine learning. The algorithm namely K-nearest neighbour algorithm is used to find the functional dependency in the given data. The concept of using data dictionary also yields effective results. The parameter namely missing rate [4] is taken into account for evaluation.

Additive least square technique with application of support vector machine which helps in performing classification of the data which are missing is presented in [3]. Cross validation strategy with ten folds is performed to correctly classify the data. The strategy is verified by measuring the accuracy factor through mean and standard deviation values [6] for the given data. The research in [20] provides embedding based calculation of the missing data through non-linear technique that bind the vector label. The proposed system is evaluated for its performance and also by the time taken for training the dataset.

The method of finding the missing data and grouping them is done through sampling in [21]. This method helps in omitting the missing data by calculating the error. Based on the accuracy and error calculation, the proposed system is evaluated. SVM based model which does not require selection of planes is investigated in[16]. The system is evaluated

TABLE I. VARIOUS EXISTING MISSING DATA HANDLING TECHNIQUES

Reference	Year	Proposed Approach	Merits	Evaluation Parameters
[1]	2016	Missing Rate Oriented selective (MROS) algorithm	Achieve effective mean Identification rate (MIR) with minimal imputation effort.	Mean Identification Rate
[2]	2018	Functional dependency based techniques, Machine learning based KNN	Functional dependency and data dictionary provide efficient results	Missing Rate
[3]	2018	Novel transfer-based additive least square support vector machine (LS-SVM)	Perform direct classification on the missing data	Ten fold cross validation strategy, Calculation of mean and SD of accuracy
[20]	2019	Embedding based method that non-linearly embeds label vector	Accurate prediction of tail labels	Prediction performance and Training Time
[21]	2014	ImputationTechnique based on sampling method	Overcomes the missing data using copulas using small error	Accuracy error
[16]	2008	SVM Kriging Technique	This model does not requires selecting variogram models	Mean error, Mean absolute error, Root mean square prediction error.
[22]	2016	BayesianNetwork, Multilayer perceptron, C.4	Higher accuracy rate is provided by KNN and optimal endurance by MLP	Accuracy and endurance

using the parameters like root mean square and absolute error [6] which helps in effective determination of the proposed technique.

Table I summarizes the various techniques for handling missing data and the merits and evaluating parameters.

III. PROPOSED METHODOLOGY

A. Panel Data

Panel data is a multidimensional-format data involving measurements over varied time. The multi-dimensional format represents the various attributes of a dataset constituting a complete dataset. Time series data also comes under the panel data. The dataset with the primary data element occurring n number of times in a particular time series makes it worth investigating. A balanced panel is one in which the panel data is continuously observed in every time interval as represented in Equation 1.

$$NO = P * T \tag{1}$$

Where, NO is the Number of Observations in the data, P denotes the panel members and T denotes the time period.Where, I (I=1....n) is the individual factor and T (T=1....t) is the time factor. At the same time, Equation 2 represent the panel data

$$P_{(IT)} \tag{2}$$

Using the panel data, the model can be constructed as shown in Equation 3 and Equation 4:

$$Q_{IT} = \mu + \sigma P_{IT} + F_{IT} \tag{3}$$

$$F_{IT} = \gamma_I + G_{IT} \tag{4}$$

Where, G_{IT} is the component which varies with time and γ_I is the specific thing to a particular member and fixed for a time interval.

B. Pre-Processing

Data in reality has a noisy and incomplete [25]. To address the preprocessing, various techniques of data cleaning integration, and reduction are incorporated to make it more consistent[26] . For the proposed technique CGA-SVM, the data cleaning process involves identification and addressing of missing values. Further data values are sorted and arranged into their respective buckets; a process called binning. Values that do not reflect any cluster are identified, differentiated through outlier detection techniques. Redundant values in data aggregation process are then removed after which the process of normalizing values is done. During data reduction, the attribute dimension which in this case is the size is reduced and data compressed making it easier to handle. The pre-processing

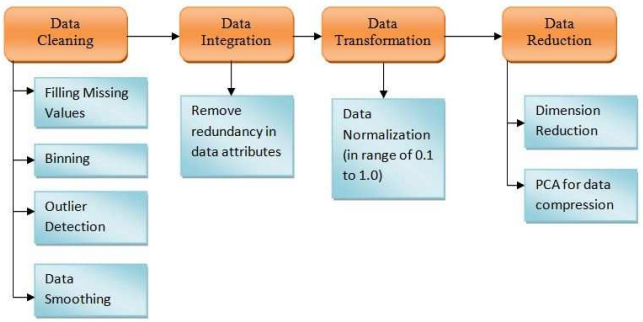

Fig. 1. Data Pre-processing Techniques for CGA-SVM Technique

technique for CGA-SVM is initially based on the data cleaning procedures as shown in Fig 1. Data cleaning has the ability to fill missing values based on mean or median values of all the values. When a missing value occurs, that individual tuple may be omitted from the general dataset. Such omission may not be efficient as key details could be missed. However, when

many columns of that tuple are missing, the omission technique would work.

Missing values can be keyed in manually but only small datasets with fewer tuples. Replacement of missing data with global constant fixed based on the relevant dataset can also apply in such a situation. Data becomes noisy by using measuring instruments that make faulty calibrations. Binning is the next technique as it helps in classifying data into several buckets. Smoothing is also a data cleaning strategy, which involves replacing bin values by either mean or with the close boundary value. The values, which do not fit any of the group, are termed as outliers and are being handled with the dataset. Therefore, the pre-processing takes place efficiently starting from cleaning and further proceeds until reduction with intermediate steps being executed.

C. Correlation

Correlation is a measure of association between two attributes and also the nature of the relationship[27]. The correlation coefficient value lies between -1 to +1. Correlation a mathematical value which describe a relationship between one or more independent variables with dependent variable. For example, a correlation can be a connection between two variables (numeric) values. If increasing happened to one variable value, then the other one also will increase (or decrease). However, potential predictive power in Correlations make it valuable: use or act on the value of one variable to predict or modify the value of the other. Furthermore, the correlation does not imply causation and a correlation does not tell us about the core cause of a relationship. The correlation method is a systematic praxis with roots going far back in human history. It is also used to analyses extremely large datasets correctly and efficiently that plays a critical role in the science of the future.

There are various correlations namely pearson, kendall and spearman which are used to measure the relationship between two attributes or variables. Pearson Correlation is a measure of the degree of relationship between linearly associated variables. The variables are required to be normally distributed. It is given as the ratio of covariance of the variables to the product of their standard deviation as shown in Equation 5.

$$\rho = \frac{cov(a,b)}{\sigma_a \sigma_b} \tag{5}$$

Where, ρ is the Pearson coefficient and a,b are the two variables.

Kendall correlation is non-parametric where there is a dependency between two variables and is represented as tau. It analyses the relationship between two variables, and provide solution between discordant and cordant pairs. Spearman correlation is ranked, depending on the variable's rank value for the operation. It is denoted as ρs.

When the three correlation techniques are adopted in both panel datasets, the outcome shows how Pearson coefficient provides better correlation values as indicated. Once the correlations between the attributes are obtained, it is subjected to the SVM classification with genetic algorithm applied.

D. CGA-SVM Technique

Data generated after analysis of the correlation factor is examined with a genetic algorithm, which is further analysed by the SVM technique. This provides room for treatment of missing values hence minimising their effects.

CGA-SVM Algorithm:

- Input: Dataset with missing values and correlation details.
- Output: Classified data with addressed missing data.

1) *Initialize each individual and then produce which is in accordance with X,i=1,2. . . .ln.*
2) *Arrive based on the signs of and*
3) *Calculate the fitness values using the fitness function for the individuals as follows: F(z)=j-f'(where f'(is the derivative of the objective function for optimization based on the correlation result and j is the constant for scaling the fitness function*
4) *If condition achieved then stop. Else move to step 5.*
5) *If the best fitness value is less than the threshold, Go to next iteration.*
6) *Do selection based on F(z). Then perform crossover between the chromosomes with same attribute values.*
7) *The Leave one out cross validation is applied as a condition for SVM with GA.*
8) *Perform new iteration of variables generation using the steps 3-7 and exit.*

In the algorithm CGA-SVM, the genetic algorithm determines the fitness value of the used variables. The fitness function is defined by determining the fitness value of variables using the objective function. it is an iteration process until the best fitness value is achieved, compared to the threshold. Mutation and crossover are also performed with the matching chromosomes. Validation is done using Leave one out technique which ensures that the proposed fitness value meets the threshold and CGA-SVM outperforms the existing techniques. The flowchart in Fig. 2 represents the correlation based Genetic algorithm and SVM combination of classification which involves parameter setting in SVM and then finding the fitness value. Once the fitness function is determined, and the required condition is satisfied, SVM model is evaluated and missing values are addressed. If condition is not met, then further genetic algorithm is being applied with the activities of selection, crossover and mutation to arrive at desired objective function.

IV. Implementation

Four benchmark panel datasets [28] are used to support the findings of this study have been selected from the UCI machine learning repository as follow:

1) Ionospher
2) Iris Plant
3) Parkinson
4) SPECTF

First correlation is applied to identify the association among the attributes in a dataset. Correlation value GA will be applied then followed by classification by SVM for meaningful data and testing of results. Table II shows the main characteristics

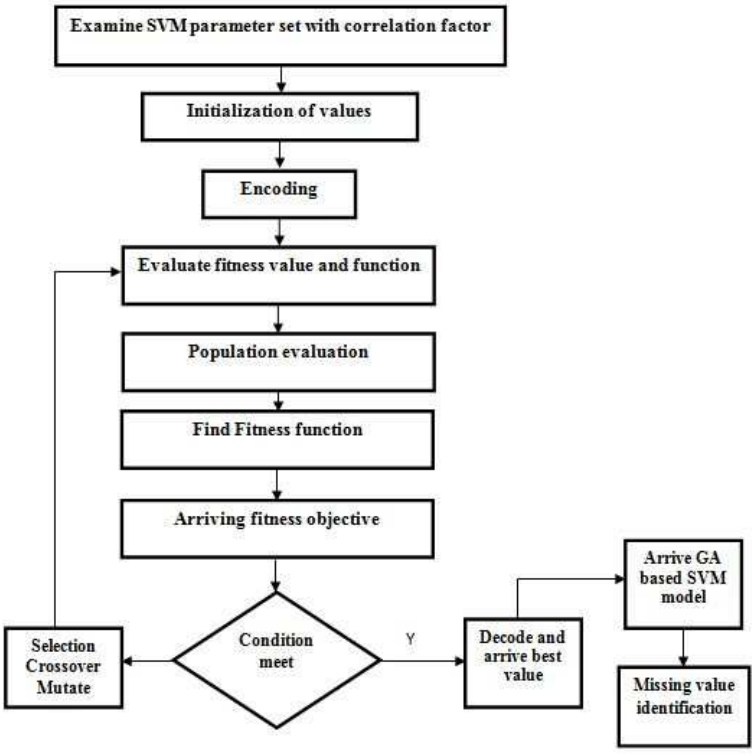

Fig. 2. Flow of CGA-SVM Approach

of each dataset with the number of instances and the number of attributes. For each dataset, the missing data with random values were used to present missing values in terms of correlation values.

TABLE II. THE MAIN CHARACTERISTICS OF EEACH DATASET

Name	Class	Instances	Features	Missing Data (Random)
Ionosphere	2	351	34	17%
Iris Plant	3	150	4	25%
Parkinson's	2	267	44	15%
SPECTF Heart	2	267	44	36%

A. Correlation Methods

Using Tidy verse package in R software, the dataset is being read and Tidy up the dataset by making every row and column clear for observation. A variable is one way to visualize the rearranged data, making the relationships between measure, class, and part a little clear. Correlation values between the various datasets parts and corresponding measures are based on what class the attributes of each data set happens to be. Reshape2 package helps to reshape benchmark dataset and establishment of variables for correlation of every measurement against the other. Pass the benchmark dataset, grouped by class, to (cor_list) function, which calculates correlations by applying the Reshape2 package. This will generate N rows as shown the Equation 6.

$$RowsNumber = (mclasses * (Nmeasurements)^2) \quad (6)$$

relation coefficients for every measurement pair. The last step of correlation is the visualization of correlations between measurements, grouped by class. Fig. 3 summarizes correlation plot box measurement according to Pearson correlation coefficients values which greater than or less than 0. We then omit 0 values indicating no relation among attribute. Fig. 4 indicates the strengths of correlation coefficients of correlation measure within four different features of used datasets.

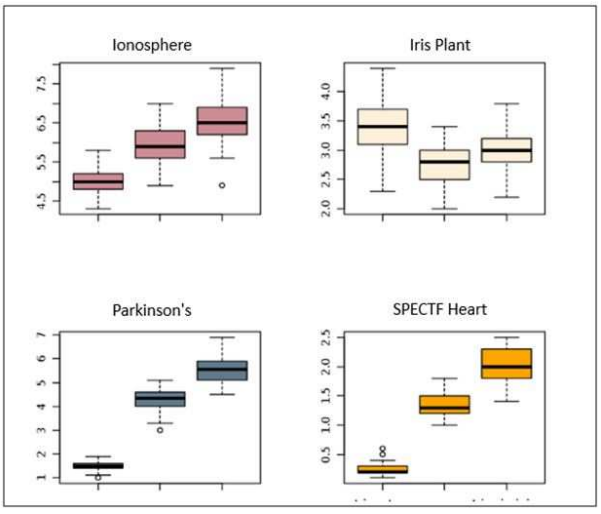

Fig. 3. Box Plot Determine the Correlation Among Different Datasets and Missing Data

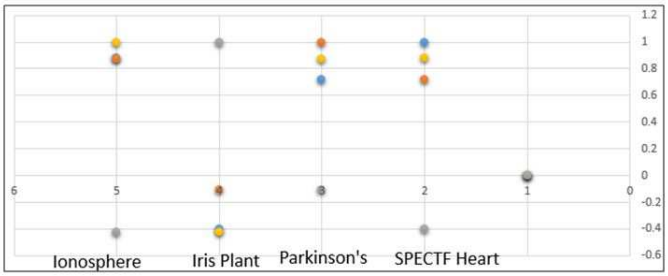

Fig. 4. Comparison of Strengths Correlation Coefficients in Correlation Measure for Four Futures within Datasets

B. Genetic Algorithm

The benefit of the proposed approach is its imputation approach based on Genetic Algorithm. Input to the algorithm is a dataset with missing values and correlation values not equating 0 and that needs imputation. Primarily, the dataset is randomly imputed. This step is relevant as it produces a temporary complete dataset for enhancements in later steps through crossover and mutation. The algorithm repeats itself continuously regressing each attribute with missing values on other attributes. Implementation of Genetic programming in the experiment adopted the Genetics package. For 100 iterations, calculate the accuracy independently running for each benchmark dataset, and the result summary is as shown in Table III and Fig. 5.

TABLE III. GENETIC ALGORITHM WITH ACCURACY VALUES FOR 21 ROUNDS

Run	Ionosphere	IrisPlant	Parkinsons
1	94.9	94.5	85.5
2	96.32	96.1	90.35
3	90	89	88.23
4	91.62	97.4	90.12
5	95.54	95.4	89.12
6	92.65	91.4	88.2
7	93.5	96.19	78.15
8	89.8	93.33	90
9	97.2	91.2	96
10	97.23	94.6	85.5
11	95.32	92.3	90.35
12	88.9	87.8	88.23
13	92.9	95.28	90.12
14	96.3	92.64	89.12
15	92.8	87.14	88.2
16	96.8	84.28	78.15
17	97.1	91.42	93.5
18	95.2	94.76	92.1
19	93.2	96.56	89
20	94.7	96.19	91.4
21	96.3	96.4	88.01
Max	97.23	97.4	96
Min	88.9	84.28	78.15
Median	94.9	94.5	89.12
Mean	94.20380	93.04238	88.54048

V. RESULTS AND DISCUSSION

The proposed system is SVM with correlation and GA applied. It is compared with the simple SVM providing accuracy and error comparison. The prediction accuracy (%) (with errors less than 10%) using SVM with correlation and GA is Accuracy1 and without correlation and GA is Accuracy2.

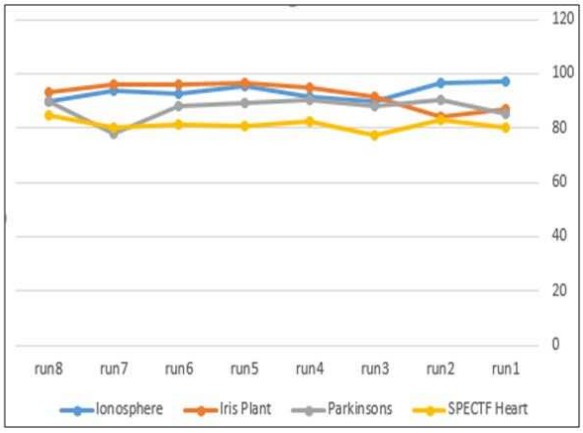

Fig. 5. GA Plot for Different Datasets

TABLE IV. ACCURACY COMPARISON OF SVM AND CGA-SVM TECHNIQUES

Dataset	Accuracy1	Accuracy2
Ionosphere	97.23	94.55
Iris Plant	96.56	95.64
Parkinson's	90.35	89.5
SPECTF Heart	84.44	80.17

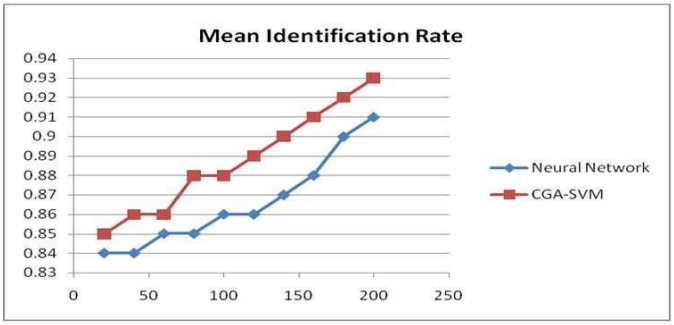

Fig. 6. Mean Identification Rate Comparison of Neural Network and CGA-SVM.

Table IV shows accuracy comparisons of various datasets for the SVM technique and CGA-SVM technique. Fig. 6 depicts the Mean Identification rate of applying the neural network approach and the proposed CGA-SVM approach. Regarding to Table IV and Fig. 6 show the best accuracy achieved in the experiments, after training with CGA-SVM. The proposed system is also compared with depicts the Mean Identification rate of applying the neural network approach by (91%) and the proposed CGA-SVM approach(93%) which mean the existing systems to handle missing values, where the results indicate a consistent accuracy hence making it better.

VI. CONCLUSIONS

Addressing missing data in the big dataset is very important. The proposed system handles the missing data through correlation technique followed by genetic algorithm imposed on support vector machine. This variant of SVM performs well as it effectively handles the missing data. The proposed

system is first subjected to correlation technique comparing the various techniques and then evaluating it using the fitness function of the genetic algorithm. The proposed CGA-SVM provides better accuracy than the existing techniques based on the standard SVM. Further, mean identification rate is used for the comparison of the proposed technique with the existing neural network approach, and the results show that the proposed technique has a higher percentage accuracy of 2% accuracy compared to existing methods.

In Future work, Grey Wolf optimization algorithm will be used to avoid the irrelevant and redundant attributes significantly, after the features are forwarded to the SVM.

REFERENCES

[1] X. Li, G. Li, and R. Fishbune, "A novel missing-rate-oriented selective algorithm for handling missing data by minimizing imputation," in *2016 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*. IEEE, 2016, pp. 234–237.

[2] I. Ezzine and L. Benhlima, "A study of handling missing data methods for big data," in *2018 IEEE 5th International Congress on Information Science and Technology (CiSt)*. IEEE, 2018, pp. 498–501.

[3] G. Wang, J. Lu, K.-S. Choi, and G. Zhang, "A transfer-based additive ls-svm classifier for handling missing data," *IEEE transactions on cybernetics*, vol. 50, no. 2, pp. 739–752, 2018.

[4] T. Yeoh, S. Zapotecas-Martínez, Y. Akimoto, H. Aguirre, and K. Tanaka, "Genetic algorithm assisted by a svm for feature selection in gait classification," in *2014 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*. IEEE, 2014, pp. 191–195.

[5] Y. Ding and J. S. Simonoff, "An investigation of missing data methods for classification trees applied to binary response data," *Journal of Machine Learning Research*, vol. 11, no. Jan, pp. 131–170, 2010.

[6] K. Sijtsma and L. A. Van der Ark, "Investigation and treatment of missing item scores in test and questionnaire data," *Multivariate Behavioral Research*, vol. 38, no. 4, pp. 505–528, 2003.

[7] D. Ugryumova, R. Pintelon, and G. Vandersteen, "Frequency response function estimation in the presence of missing output data," *IEEE Transactions on Instrumentation and Measurement*, vol. 64, no. 2, pp. 541–553, 2014.

[8] J. Li, B. Zhang, and J. Shi, "Combining a genetic algorithm and support vector machine to study the factors influencing co2 emissions in beijing with scenario analysis," *Energies*, vol. 10, no. 10, p. 1520, 2017.

[9] D. Nithya, V. Suganya, and R. S. I. Mary, "Feature selection using integer and binary coded genetic algorithm to improve the performance of svm classifier," *Journal of Computer Applications (JCA)*, vol. 6, no. 3, p. 2013, 2013.

[10] P. Stoica, J. Li, J. Ling, and Y. Cheng, "Missing data recovery via a nonparametric iterative adaptive approach," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 3369–3372.

[11] S. Lessmann, R. Stahlbock, and S. F. Crone, "Genetic algorithms for support vector machine model selection," in *The 2006 IEEE International Joint Conference on Neural Network Proceedings*. IEEE, 2006, pp. 3063–3069.

[12] S. Di Martino, F. Ferrucci, C. Gravino, and F. Sarro, "A genetic algorithm to configure support vector machines for predicting fault-prone components," in *International conference on product focused software process improvement*. Springer, 2011, pp. 247–261.

[13] L. Garg, J. Dauwels, A. Earnest, and K. P. Leong, "Tensor-based methods for handling missing data in quality-of-life questionnaires," *IEEE journal of biomedical and health informatics*, vol. 18, no. 5, pp. 1571–1580, 2013.

[14] M. Li, X. Zhou, X. Wang, and B. Wu, "Genetic algorithm optimized svm in object-based classification of quickbird imagery," in *Proceedings 2011 IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services*. IEEE, 2011, pp. 348–352.

[15] H.-B. Liu and Y.-B. Jiao, "Application of genetic algorithm-support vector machine (ga-svm) for damage identification of bridge," *International Journal of Computational Intelligence and Applications*, vol. 10, no. 04, pp. 383–397, 2011.

[16] W. Huizan, Z. Ren, L. Kefeng, L. Wei, W. Guihua, and L. Ning, "Improved kriging interpolation based on support vector machine and its application in oceanic missing data recovery," in *2008 International Conference on Computer Science and Software Engineering*, vol. 4. IEEE, 2008, pp. 726–729.

[17] S. Moridpour, T. Anwar, M. T. Sadat, and E. Mazloumi, "A genetic algorithm-based support vector machine for bus travel time prediction," in *2015 International Conference on Transportation Information and Safety (ICTIS)*. IEEE, 2015, pp. 264–270.

[18] G. Wang, Z. Deng, and K.-S. Choi, "Tackling missing data in community health studies using additive ls-svm classifier," *IEEE journal of biomedical and health informatics*, vol. 22, no. 2, pp. 579–587, 2016.

[19] W. Shi, Y. Zhu, S. Y. Philip, T. Huang, C. Wang, Y. Mao, and Y. Chen, "Temporal dynamic matrix factorization for missing data prediction in large scale coevolving time series," *IEEE Access*, vol. 4, pp. 6719–6732, 2016.

[20] A. H. Akbarnejad and M. S. Baghshah, "An efficient semi-supervised multi-label classifier capable of handling missing labels," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 2, pp. 229–242, 2018.

[21] R. Houari, A. Bounceur, A. K. Tari, and M. T. Kecha, "Handling missing data problems with sampling methods," in *2014 International Conference on Advanced Networking Distributed Systems and Applications*. IEEE, 2014, pp. 99–104.

[22] O. M. Prabowo, K. Mutijarsa, and S. H. Supangkat, "Missing data handling using machine learning for human activity recognition on mobile device," in *2016 International Conference on ICT For Smart Society (ICISS)*. IEEE, 2016, pp. 59–62.

[23] K. Pelckmans, J. De Brabanter, J. A. Suykens, and B. De Moor, "Handling missing values in support vector machine classifiers," *Neural Networks*, vol. 18, no. 5-6, pp. 684–692, 2005.

[24] D. P. Mesquita, J. P. Gomes, F. Corona, A. H. S. Junior, and J. S. Nobre, "Gaussian kernels for incomplete data," *Applied Soft Computing*, vol. 77, pp. 356–365, 2019.

[25] W. J. Alzyadat, A. AlHroob, I. H. Almukahel, and R. Atan, "Fuzzy map approach for accruing velocity of big data," *Compusoft*, vol. 8, no. 4, pp. 3112–3116, 2019.

[26] A. Alhroob, W. J. Alzyadat, I. H. Almukahel, and G. M. Jaradat, "Adaptive fuzzy map approach for accruing velocity of big data relies on fireflies algorithm for decentralized decision making," *IEEE Access*, vol. 8, no. 1, pp. 2169–3536, 2020.

[27] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise reduction in speech processing*. Springer, 2009, pp. 1–4.

[28] C. Dua, Dheeru & Graff, "Uci machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml