IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 12, 2021, accepted June 30, 2021, date of publication July 6, 2021, date of current version July 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3095240

Machine Learning Analysis for Data Incompleteness (MADI): Analyzing the Data Completeness of Patient Records Using a Random Variable Approach to Predict the Incompleteness of Electronic Health Records

VARADRAJ P. GURUPUR[1], (Senior Member, IEEE), AND MUHAMMED SHELLEH[2], (Member, IEEE)
[1]Department of Health Management and Informatics, University of Central Florida, Orlando, FL 32826 USA
[2]Department of Computer Science, University of Central Florida, Orlando, FL 32826 USA

Corresponding author: Varadraj P. Gurupur (varadraj.gurupur@ucf.edu)

This work was supported by the Office of Faculty Excellence, University of Central Florida.

ABSTRACT The purpose of this article is to propose a methodology involving various methods that can be used to predict the data incompleteness of a dataset. Here the investigators have presented data incompleteness as both continuous and discrete random variables. In addition the investigators used transfer entropy for the purpose of advancing the science associated with the analysis of data incompleteness of electronic health records. The underlying methodology has been coined as ''Machine Learning Analysis for Data Incompleteness'' (MADI) with the intention of developing a possible solution to data incompleteness in electronic health records. MADI advances the analysis of data incompleteness with the use of Kolomogorov Smirnov goodness of fit, mielke distribution, and beta distributions for a holistic analysis. Alongside the methodology presented, the investigators explored stochastic gradient descent, generalized additive models, and support vector machines for comparison. Overall, the investigators have presented a complete set of methods and algorithms to help predict data incompleteness in a medical setting and provided suggestions for practical applications into the prediction of data incompleteness.

INDEX TERMS Health informatics, big data models, data completeness, probability density, Kolomogorov-Smirnov test, support vector machine, stochastic gradient descent, generalized additive model, electronic health records.

I. INTRODUCTION

In the world of medical informatics, the critical importance of the data completeness of electronic health records is becoming more and more clear. As the digital transformation of medical facilities across the world expands and provides greater computational abilities to health professionals, patient data has been viewed closer than ever before. Whether the data need exists for research, hospital demographics, or even for the patients themselves, the need for completeness of each entry in a medical records system increases. In the past Nasir *et al.* [1], have introduced the idea of using an algorithmic approach towards solving this problem. However, given the advances in artificial intelligence and machine learning there is a critical need for a more advanced approach that can be effectively used to predict data incompleteness in electronic health records. For this purpose, the investigators involved in the project delineated in this article describe the use of advanced machine learning methods to predict this data incompleteness. Specifically, the core objectives of the work delineated in this article are as follows:

- advance the application of probability distributions to improve the prediction of data incompleteness in electronic health records, and
- advance the application of transfer entropy and ontologies in predicting the same.

For this purpose, the investigators present a methodology termed as Machine Learning Analysis for Data Incompleteness (MADI).

The associate editor coordinating the review of this manuscript and approving it for publication was Adnan Kavak.

II. BACKGROUND

As data warehouses within medical and hospital systems increase, the representation and scope of the data within electronic medical systems has increased as well. Given the increase in scope and patient population representation, data availability and completeness presents itself as a much more critical issue within the electronic health system used. When a doctor or provider has a patient present, data will be entered to represent the visit and what the situation is involving the patient. However, in many different cases, the data is entered quickly or hastily, and some of the data turns incomplete or blank, as a result [2]. This is a major issue since the need for completeness in clinical or medical data is a must when working in medical research or even when trying to backreference past patient data.

III. METHODS

Based on the argument presented on data completeness, the investigators provide a comprehensive study in analyzing data incompleteness of electronic health records. Here the investigators use Probability Density Function (PDF) as a representation of patient record's incompleteness, where the data incompleteness is perceived as a random variable [3]. In probability and statistics, a random variable is described informally as a variable whose values depend on the outcomes of a random phenomenon [4].

To propose our algorithm, we define the following variables:

- Completeness Parameter Variable (CPV): where x_{wz} is the measure of completeness for the data field located in w^{th} row and z^{th} column. Measured in a binary method, 1 represents a complete data field and 0 represents an incomplete field. [5]
- Completeness Scoring Variable (CSV): Where, CSV_z is the completeness measure of column z and its value is between 0 and 1:

$$\text{CSV}_z = \frac{r_{w=1} x_{wz}}{r} \tag{1}$$

  where r is number of rows
  $1 \leq z \leq$ number of columns (c)[1]
- DIM_z: defines the Data Incompleteness Measure of column z (or the Incompleteness Ratio of column z):

$$0 \leq \text{DIM}_z = 1 - \text{CSV}_z \leq 1 \tag{2}$$

According to Algorithm 1, the investigators compute DIM for each column of the experimental dataset. This is followed by the generation of the histogram of the entire dataset [6], based on the incompleteness ratio of each column [7].

A. DISTRIBUTION FITTING

With regards to predicting the incompleteness of each dataset presenting, one of the most critical tasks in data pre-processing is distribution fitting. To begin the process of distribution fitting, we have to understand the parameters of each dataset we're working with, and provide a general idea of how each model will work, given those parameters [8]. Using the SciPy package, we can call a Maximum Likelihood Estimator (*MLE*) for parameter estimation on each of the datasets used.

[1]In practice, the measure $CSV_1 = \mathbf{1}$, since the first column usually contains the header of each row, and therefore remains complete.

Algorithm 1: Plotting the Histogram for the Experimental Dataset

```
procedure Plot the Histogram()
Initialize DIM_data = 0
for 1 ≤ z do
    Compute DIM_z
    DIM_data[z] ← DIM_z
end
Initialize the histogram bins
Plot the corresponding histogram
```

The Kolomogorov Smirnov test [9] is used to determine if a sample distribution comes from a specific distribution. It is based on the empirical distribution function (*ECDF*) [10]. Given $N_{samples}$ number of ordered data samples Y_1, Y_2,…, $Y_{Nsamples}$, this test is defined by [11]: i) the data which fit a specified distribution, ii) the data which do not fit the specified distribution, iii) Test Statistic D_{KS}: (3)

$$D_{KS} \max_{1 \leq w \leq N_{samples}} (F(Y_w) - \frac{w-1}{N_{samples}}, \frac{w}{N_{samples}} - F(Y_u)) \tag{3}$$

Distribution fitting is not the only necessary step required in data engineering, but rather, one of two steps. The second critical step in finding a best fit distribution is testing the proposed model. The method we used in our experimentation was the Kolomogorov − Smirnov test, as seen previously in Section III-A1. The Kolomogorov Smirnov test has been widely used in a variety of statistical models, as seen in studies such as [12] and [13]. As a result, we can use the Kolomogorov Smirnov test as a method to finding the best fitted distribution. As for a few assumptions of the model, the data is assumed to be standardized and each distribution (in this case, the 88 distributions known and available within the SciPy library) must be applied to the dataset.

Algorithm 2 represents our method to finding the best distribution to fit to the outputted histogram.

B. COMPLETENESS GRAPHING

Aside from fitting the data with a proper distribution, measures of incompleteness are much easier to grasp and utilize when viewing them in a more tangible format. To begin data post-processing and representation, it is critical to utilize all data available, along with the chosen distribution seen in Section III-A1. We utilized the MissingNo Library [14], [15] to produce high-quality depictions of the data represented throughout experimentation. Algorithm 3 presents our pseudo-code for creating the visualizations following distribution fitting.

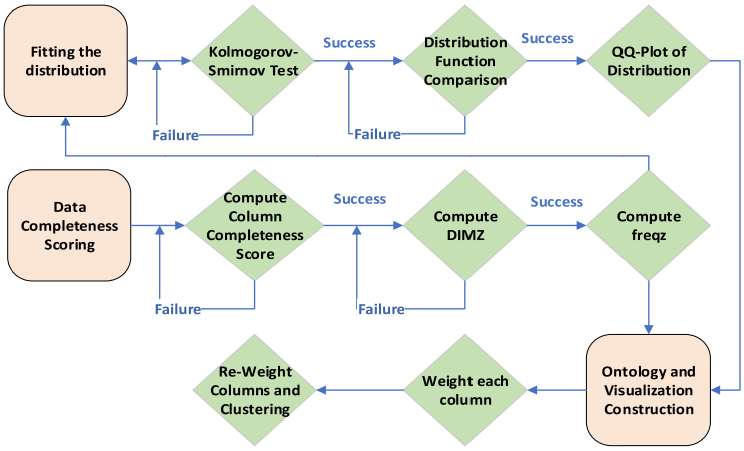

FIGURE 1. Algorithm overview.

Algorithm 2: Finding the Best Fit for the Histogram

procedure Fit the best distribution to the histogram()
total distributions within the SciPy library = 88
for *each distinct* DIM_z **do**
  Compute its frequency freq_z
  Let $\text{freq}_{data}[z] \leftarrow \text{freq}_z$
end
Run the distribution fitting algorithm on freq_{data}
for $1 \leq w$ *88* **do**
  if *distribution[w] is the best match* **then**
    Let *TheBestFit distribution*$[w]$
  end
end
Apply the Kolomogorov Smirnov test to measure the goodness of the fit

Algorithm 3: Representing the Fitted Data for Each Dataset

procedure Present data from *TheBestFit* and process for each diagram, along column max = 50
for *each set* DIM_z *represented* **do**
  Take its fit for $\text{freq}_{data} \leftarrow x_{wz}$
  Let freq_{data} represent the grid layout
end
Set the representation output algorithm for $fr_{data}[z]$
for $1 \leq z$ 50 **do**
  if *x[wz] is presented* **then**
    Let *ImageTarget* $\leftarrow x[wz]$
  end
end
Output the ImageTarget for each presented dataset

1) INTERSECTION CALCULATION

As mentioned above, we utilized the MissingNo Library [14] as one of our main visualization tools in terms of converting the data from logistic and applied regression models into visualized representations of how the dataset looks.

However, another important aspect behind designing the ontology and improving the overall strength in the representation of information entropy within the dataset is the design and calculation of intersections within the entropic points of the data [16]. As seen in previous works, information entropy presents itself in different contexts, and can be represented using comparison or coordination. Using the available modules within SciPy and MissingNo, a method can be reached where the different levels of entropy in the ontological map are represented and give a greater context to how the incompleteness of the data connects at different points within the presented structural aspects of the data itself [17]. Another key aspect when it comes to measuring the entropy of the datasets is seen in the information contents of a node itself within the hierarchical structure created [18]. Here, a probabilistic generalization is

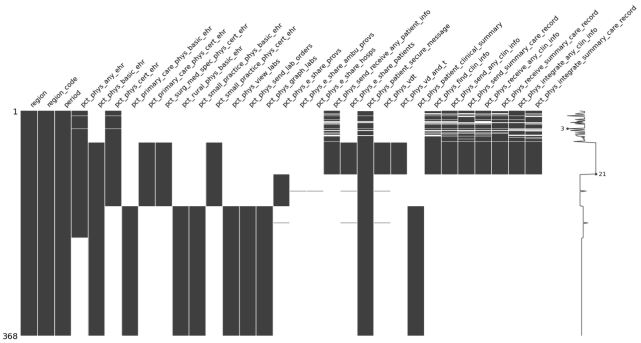

FIGURE 2. Data matrix of incompleteness following Algorithm 4.

created for each data point within the set, and the lengths of each node is generated based on the probability of a given node [column]. As nodes in the hierarchical structure increase, the overall information representation increases in tandem [19].

Algorithm 4 presents our method for producing the intersections found in our ontological representation of the mixed data-type dataset, as seen in 6.

Algorithm 4: Intersection Calculation and Entropy Information Representation

procedure Present intersections for each representation of x
for *For all numerical columns present for* $x[i]$ **do**
  Weight the columns present in $x[i]$ for $w[i] = 1$, for each non-null column
  while x *is uncategorized* **do**
    Output a set of values y, where $y = y, -y + 1, \ldots, -1, 0, 1, \ldots, y - 1, y$, where y is a measured overlap parameter for all columns x
    for *Each Non-Null set of y (represented by e)* **do**
      Process each column where overlap occurs by a measure $x[\boldsymbol{j}] = x[i] + \frac{e}{f}$, where f is to be used to divide the column integer overlaps
      Re-weight each column of the set x by $w[\boldsymbol{j}] = 1 - \frac{e}{y+1}$, where the final fraction will be raised to a measured coefficient power
    end
  end
end

IV. EXPERIMENTATION AND RESULTS

In this section, the authors illustrate the results obtained after implementing Algorithm 1, Algorithm 2, Algorithm 3, and Algorithm 4 using SciPy, NumPy, and MissingNo, which have been discussed in Section III and in Section III-B1. To evaluate our proposed algorithms, we used three different types of datasets, as follows; integer-based, mixed-data, and string- based datasets [20]. As experimentation continued, one of the most surprising findings seen was in the ways measuring data incompleteness worked when testing different models.

When it comes to traditional and probabilistic statistical methods, the distribution and representation of the data held true in majority of the models tested. Among the most useful models were the Kolomogorov Smirnov test, mielke representation, logistic regression, and other non-parametric representations of the distribution.

As seen in Algorithm 1, Algorithm 2, Algorithm 3, and Algorithm 4, the use of seemingly simple statistical methods can help mitigate much needed computation when dealing with how data columns are complete and re-applying the findings into other contexts. Starting with Algorithm 1, the whole of the input data is used to output a histogram, and a data incompleteness measure is applied to each column of the inputted medical data to provide an idea of how the data is distributed throughout the dataset.

In Algorithm 2, the histogram produced in Algorithm 1 is then re-applied and utilized as a method to form a line or match of best fit, given the present output data. A maximum likelihood estimator is used, along with distribution fitting methods, to output a best-fit model. After the algorithm procedure is finished, the Kolomogorov Smirnov test is then used to provide a tangible visualizer on how well the fit actually is.

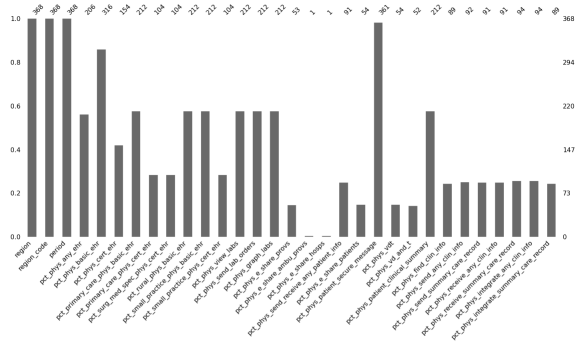

FIGURE 3. Bar plot of incompleteness following Algorithm 4.

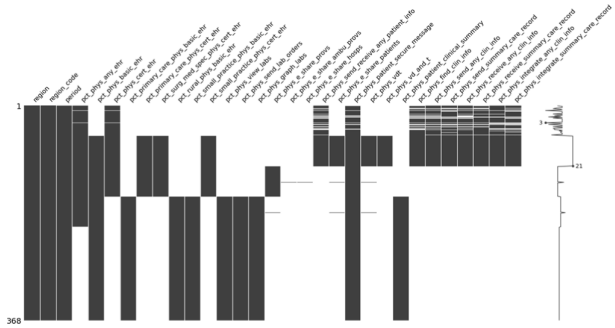

FIGURE 4. Correlation matrix of incompleteness following Algorithm 4.

In Algorithm 3, the data represented in Algorithm 1 and Algorithm 2 is then utilized as a method to re-introduce the findings in a different context. Using the incompleteness and best fit measures, the data can be reapplied and visualized in a series of different plots. The experimentation lead to a series of plots being seen as ideal, which includes a heatmap, ontology, and bar graph. The data is presented in easy to read formats, and provides a much better look on how the dataset is distributed.

Finally, in Algorithm 4, a method for how the ontology is developed is shown. Given the data from Algorithm 1, the algorithm can weight each column and classify each into it's own separate block. Entropy weights and distances are determined, which allows the data to be shown in the context of information entropy.

V. ALTERNATIVE METHODS TESTED

In this section, the investigators present the results of experimentation when applying the proposed algorithm under different contexts. These include Stochastic Gradient Descent, Generalized Additive Models, and Support Vector Machines. While these may not have been used in the final proposed solution, the results found do provide an interesting look into how data incompleteness can be approaching with modern machine learning techniques.

A. STOCHASTIC GRADIENT DESCENT

With regards to stochastic gradient descent, past research suggests that optimization of objective functions may improve upon total measures of data incompleteness and improve

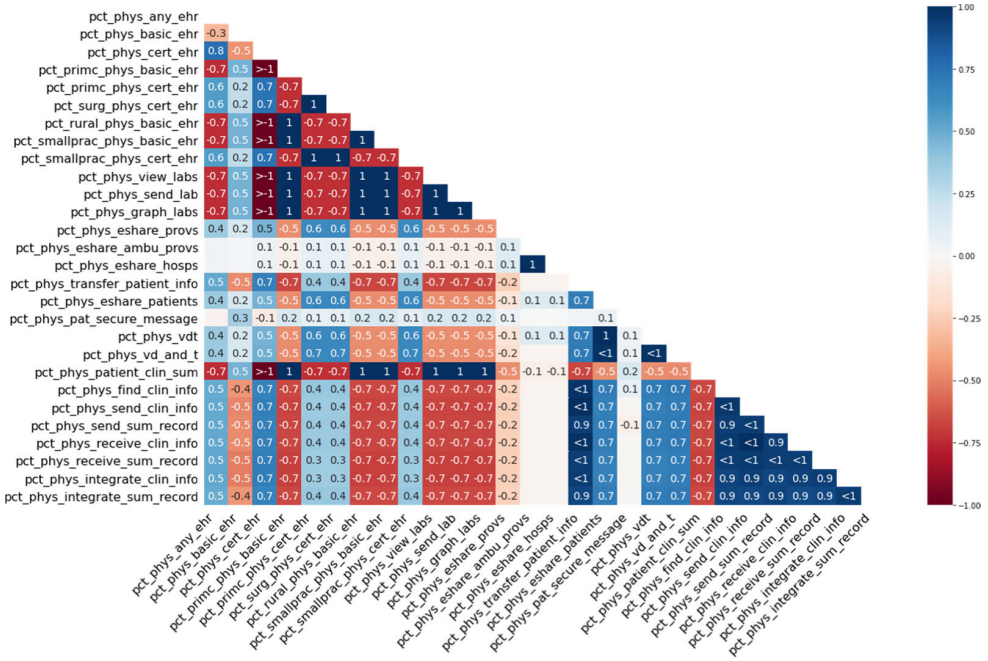

FIGURE 5. Heatmap of data incompleteness: For a binary measure, 1 represents complete data (in a correlative analysis), and −1 represents always incomplete data.

upon the general output of data features, following data engineering. However, in practice, a few issues arose when attempting to implement stochastic gradient descent or even regular gradient descent models. The first issue seen in experimentation was with regards to the objective function and optimization of it. While there is a general ''fit'' for the data, it cannot be seen given that the data is missing. Past researchers have suggested filling in the missing data to allow for the objective function to be optimized, however, given the goal of finding a measure of incompleteness, is an incompatible suggestion. Another issue is seen with how data presents itself when incomplete. When data is incomplete, measures of optimization and general improvements will not affect the overall fit of the data, and therefore shouldn't be used when the goal is to view the data as raw as possible (or as it was originally inputted.

B. GAMS, SVMS, AND UNSUPERVISED MACHINE LEARNING MODELS

The next models attempted during testing were Generalized Additive Models, Support Vector Machines, and some unsupervised methods. One of the main questions raised was in how several methods could produce the same result, and the answer shown in testing was unfortunately disappointing. To begin with Generalized Additive Models, past studies have used the term non-ignorability and suggested a sensitivity analysis when utilizing the models on incomplete data. In regards to Support Vector Machines, suggestions turned to the direction of adjusting classifiers and preparing for ''missing'' bias. In the final testing rounds, unsupervised models pointed towards trying out similarity scores and preparing the study using a-priori methods.

The reason these were grouped together in a section was due to the end result; over-fitting of the datasets and overall lack of results was displayed in each of these methods. Similar issues occurred through each iteration of testing; the incompleteness of the dataset would cause the model to act inappropriately, or the resulting output from each of the models showed filtering when it comes to measuring and testing the missing data. As a result, each of these methods were not viable solutions when it came to attempting to create methods surrounding them.

However, past research does suggest that further testing could lead to an improved model, especially when it comes to unsupervised learning. Reapplying the methods seen in noisy dataset research, frequency-inverse analysis methods seem to have the greatest viability with regards to measuring the incompleteness [21] of a medical record system. The biggest hurdle for future research seems to come from model training and large-scale data warehousing, as often seen in the medical networks of a typical hospital or medical complex.

VI. DISCUSSION

Over time, the importance and overall need for data completeness measurements within the medical and allied health professions have increased and provided a clear look into how critical systems can fail without such parameters in mind. Solutions and methods to addressing the lack of algorithmic approaches to data completeness have slowly trickled in from different parts of academia, but there are very few papers or available methods giving a robust and finite solution in a medical setting. Given the need and lack of such an application, adding in data completeness operations into medical records systems would allow for overall medical warehouses

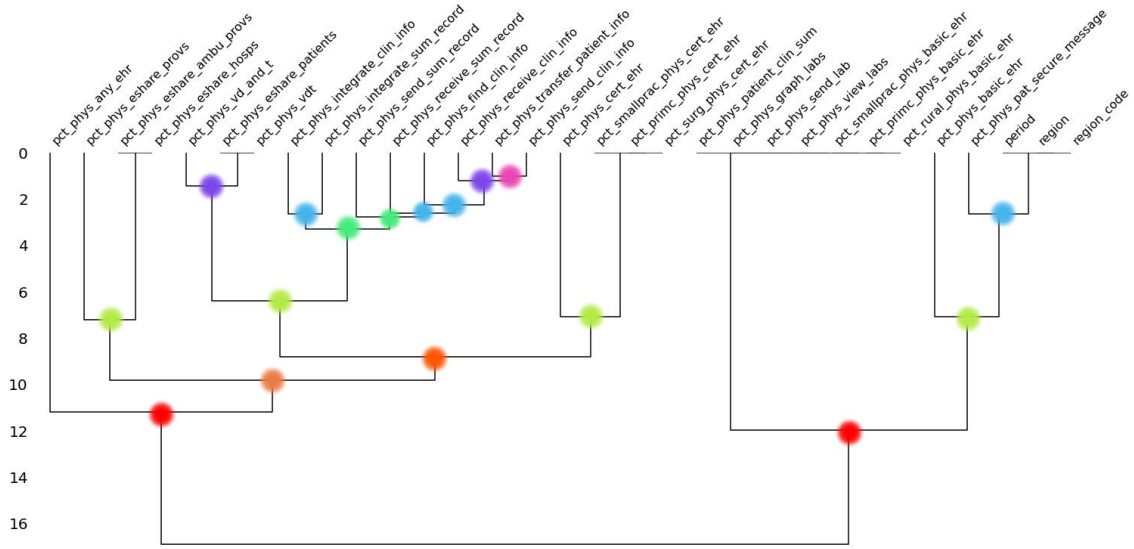

FIGURE 6. Ontology of the information entropy associated with data incompleteness.

to get insight onto where their physicians and practitioners are missing information, and give a better spread of the clinical data available not only to the offices practicing, but to the patient populations being served.

Throughout the experimentation and overall research process, the proposed ideology has centered on finding a complete solution to solving the issues seen in data availability and completeness in medical records systems. Utilizing novel algorithmic and statistical methods, we wanted to see how several regular, robust, and noisy methods could address the problem, along with some optimization and regression techniques [22]. Using a combined method involving several of these techniques, our algorithmic system focuses on a statistical calculation of the overall columns seen in a given dataset (leading to completeness scoring and a plotted histogram for visualization), followed by the application of several different statistical spread methods to present a best-fit representation for the overall completeness. The fitting of the spread is then verified using a Kolomogorov − Smirnov test [23].

Alongside the proposed solution, there are several other methods the investigators have tested to attempt to see their effects on overall data completeness and whether or not they worked better than the proposed algorithmic solution. Such examples include stochastic gradient descent, generalized additive models, and support vector machines. The main issue we ran into was over-fitting and the inability to work as an algorithmic method on several different medical datasets the methods were tested on. As a result, further experimentation was abandoned in favor of a statistical and optimization focused method.

A. COMPARISON WITH OTHER RELATED WORK

With regards to other previous works, the researchers sought to explore novel and computational methods to apply data incompleteness to an applied health dataset.

Some of the studies done in the past relied heavily on supervised machine learning methods [24], [25], or used differing statistical methodologies [26]–[29]. Our focus in this study wanted to provide two clear goals:

1) Provide a practical utilization of machine learning analysis with regards to electronic health record data incompleteness.
2) Propose implementation and software development patterns for usage in large-scale medical networks.

In an effort to provide the best solution moving forwards, the main solution revolved around optimization and implementation in a warehouse system. Given these guidelines and ideologies at play, the comparison with past works reveals that there is some room for improvement in medical records systems. Majority of currently available algorithms and modules, the solutions typically involve using smoothing or data collection models, which can have issues in the long run [30]. Overall, the uniqueness of the work delineated in this article presents an integration of information entropy, probability distributions, and ontologies to the problem of data incompleteness.

B. LIMITATIONS

Current limitations in the proposed methods appear when datasets are not a large enough to be supported by the available supervised learning methods or by the algorithmic approach over the four algorithms. This occurs in several different use cases, such as when the data cannot be converted into an ordinal format, or when the dataset is too small to be split and calculated through regression or statistical analysis.

The best example of this comes in the usage of a smaller module dataset within a medical EHR system. When working in modules and certain units, you may be limited to a small set of time-series data. Such is seen when you want to pull medical records from a certain unit within a defined time period. When this occurs, the column and row length will limit the

algorithmic calculations on how much incompleteness can be transmitted and presented as a method of entropy. A further consequence is that, even though the measure within the data space remains binary, the overall representation and spread of the data can become very uneven and unclear over shorter datasets.

Another key limitation of this research is that the investigators have not taken into considerations of missing at random, missing completely at random, and not missing at random [31], [32]. The involvement of these factors will involve questions such as how data was gathered in the publicly available datasets used for experimentation. Here the investigators will have to accept with all honesty that they do not have information on these factors and thereby not being able to make suitable comments on this matter.

The last limitation seen within the algorithmic method is in how the system works. Through the series of algorithms, the spread and overall coverage of the data measurements are limited to the completeness and relevant information of the columns in a given dataset. Other information related to the dataset will either have to be analyzed separately or reconciled as a different algorithm.

VII. CONCLUSION

To summarize, the investigators have described an experimentation that provides the scientific community a new horizon on the application of probability distributions, transfer entropy, ontologies to the problem of analyzing data incompleteness; thereby, advancing the science of medical informatics. Specifically, the core contributions of MADI are as follows:

- Advancing the science of transfer entropy applied to the problem of data incompleteness.
- Advancing the application of ontologies in analyzing electronic health records.
- Advancing the application of probability distributions to advance machine learning applied to data incompleteness of electronic health records.

Furthermore, the article also presents some insights into Support Vector Machines, stochastic gradient descent, and generalized additive models with respect to this problem. In future, the investigators plan to advance MADI with more advanced machine learning approaches [33].

REFERENCES

[1] A. Nasir, V. Gurupur, and X. Liu, "A new paradigm to analyze data completeness of patient data," *Appl. Clin. Informat.*, vol. 7, no. 3, p. 745, 2016.

[2] H. Estiri, J. G. Klann, S. R. Weiler, E. Alema-Mensah, R. J. Applegate, G. Lozinski, N. Patibandla, K. Wei, W. G. Adams, M. D. Natter, E. O. Ofili, B. Ostasiewski, A. Quarshie, G. E. Rosenthal, E. V. Bernstam, K. D. Mandl, and S. N. Murphy, "A federated EHR network data completeness tracking system," *J. Amer. Med. Inform. Assoc.*, vol. 26, no. 7, pp. 637–645, Jul. 2019.

[3] R. Nabi, R. Bhattacharya, and I. Shpitser, "Full law identification in graphical models of missing data: Completeness results," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 7153–7163.

[4] C. Chatfield and H. Xing, *The Analysis of Time Series: An Introduction With R*. Boca Raton, FL, USA: CRC Press, 2019.

[5] M. Reinikainen, P. Mussalo, S. Hovilehto, A. Uusaro, T. Varpula, A. Kari, V. Pettilä, and Finnish Intensive Care Consortium, "Association of automated data collection and data completeness with outcomes of intensive care. A new customised model for outcome prediction," *Acta Anaesthesiologica Scandinavica*, vol. 56, no. 9, pp. 1114–1122, Oct. 2012.

[6] A. A. Verma, S. V. Pasricha, H. Y. Jung, V. Kushnir, D. Y. F. Mak, R. Koppula, Y. Guo, J. L. Kwan, L. Lapointe-Shaw, S. Rawal, T. Tang, A. Weinerman, and F. Razak, "Assessing the quality of clinical and administrative data extracted from hospitals: The general medicine inpatient initiative (GEMINI) experience," *J. Amer. Med. Inform. Assoc.*, vol. 28, no. 3, pp. 578–587, Mar. 2021.

[7] G. Erhuanga, "The relationship between the extent to which physicians use key electronic health record (EHR) interoperability domains and their performance on quality measures over time," Ph.D. dissertation, School Health Professions, Rutgers Univ., New Brunswick, NJ, USA, 2020.

[8] K. I. Park and Park, *Fundamentals of Probability and Stochastic Processes With Applications to Communications*. Cham, Switzerland: Springer, 2018.

[9] F. J. Massey, "The Kolmogorov-Smirnov test for goodness of fit," *J. Amer. Stat. Assoc.*, vol. 46, no. 253, pp. 68–78, Mar. 1951.

[10] R. Castro, *The Empirical Distribution Function and the Histogram* (Lecture Notes 2WS17-Advanced Statistics). Eindhoven, The Netherlands: Eindhoven Univ. of Technology, Department of Mathematics, 2015, vol. 4.

[11] M. Karson, "Handbook of methods of applied statistics. Volume I: Techniques of computation descriptive methods, and statistical inference. Volume II: Planning of surveys and experiments. I. M. Chakravarti, R. G. Laha, and J. Roy, New York, John Wiley; 1967," *J. Amer. Stat. Assoc.*, vol. 63, no. 323, pp. 1047–1049, 1967.

[12] R. H. C. Lopes, P. R. Hobson, and I. D. Reid, "Computationally efficient algorithms for the two-dimensional Kolmogorov–Smirnov test," *J. Phys., Conf. Ser.*, vol. 119, no. 4, Jul. 2008, Art. no. 042019.

[13] I. G. Abrahamson, "Exact bahadur efficiencies for the Kolmogorov-Smirnov and Kuiper one- and two-sample statistics," *Ann. Math. Statist.*, vol. 38, no. 5, pp. 1475–1490, Oct. 1967.

[14] A. Bilogur, "Missingno: A missing data visualization suite," *J. Open Source Softw.*, vol. 3, no. 22, p. 547, Feb. 2018.

[15] A. K. Tripathi, G. Rathee, and H. Saini, "Taxonomy of missing data along with their handling methods," in *Proc. 5th Int. Conf. Image Inf. Process. (ICIIP)*, Nov. 2019, pp. 463–468.

[16] R. Lukyanenko, J. Parsons, and Y. Wiersma, "The impact of conceptual modeling on dataset completeness: A field experiment," in *Proc. Int. Conf. Inf. Syst. (ICIS)*, Auckland, New Zealand, 2014, doi: 10.13140/2.1.4852.6408.

[17] A. Sadeghi, C. Lange, M.-E. Vidal, and S. Auer, "Integration of scholarly communication metadata using knowledge graphs," in *Proc. Int. Conf. Theory Pract. Digit. Libraries*. Athens, Greece: Springer. 2017, pp. 328–341.

[18] L. Barnett, A. B. Barrett, and A. K. Seth, "Granger causality and transfer entropy are equivalent for Gaussian variables," *Phys. Rev. Lett.*, vol. 103, no. 23, Dec. 2009, Art. no. 238701.

[19] M. Cho, C. Choi, W. Kim, J. Park, P. K. A. Canedo, B. Abderazek, and M. Sowa, "Comparing ontologies using entropy," in *Proc. Int. Conf. Converg. Inf. Technol. (ICCIT)*, 2007, pp. 873–876.

[20] N. T. Longford, "Describing incompleteness," in *Missing Data and Small-Area Estimation: Modern Analytical Equipment for the Survey Statistician*. 2005, pp. 19–35.

[21] Y. van Gennip, B. Hunter, A. Ma, D. Moyer, R. de Vera, and A. L. Bertozzi, "Unsupervised record matching with noisy and incomplete data," *Int. J. Data Sci. Anal.*, vol. 6, no. 2, pp. 109–129, Sep. 2018.

[22] C. Holden, L. Thiamwong, D. Martin, K. M. Mathieson, and G. M. Nehrenz, "The electronic health record system and hospital length of stay in patients admitted with hip fracture," *Amer. J. Res. Nursing*, vol. 1, pp. 1–5, Jul. 2015.

[23] B. Yu, Z. He, A. Xing, and M. L. A. Lustria, "An informatics framework to assess consumer health language complexity differences: Proof-of-concept study," *J. Medical Internet Res.*, vol. 22, May 2020, Art. no. e16795.

[24] B. Hernandez, P. Herrero, T. M. Rawson, L. S. P. Moore, B. Evans, C. Toumazou, A. H. Holmes, and P. Georgiou, "Supervised learning for infection risk inference using pathology data," *BMC Med. Informat. Decis. Making*, vol. 17, no. 1, pp. 1–12, Dec. 2017.

[25] S. Juddoo and C. George, "A qualitative assessment of machine learning support for detecting data completeness and accuracy issues to improve data analytics in big data for the healthcare industry," in *Proc. 3rd Int. Conf. Emerg. Trends Electr., Electron. Commun. Eng. (ELECOM)*, Nov. 2020, pp. 58–66.

[26] N. G. Weiskopf, G. Hripcsak, S. Swaminathan, and C. Weng, "Defining and measuring completeness of electronic health records for secondary use," *J. Biomed. Informat.*, vol. 46, no. 5, pp. 830–836, Oct. 2013.
[27] M. M. Haider, K. Mahmud, H. Blencowe, T. Ahmed, J. Akuze, S. Cousens, N. Delwar, A. B. Fisker, V. P. Hardy, S. T. Hasan, and M. A. Imam, "Gestational age data completeness, quality and validity in population-based surveys: EN-INDEPTH study," *Population Health Metrics*, vol. 19, no. S1, pp. 1–18, Feb. 2021.
[28] N. G. Wysham, S. P. Wolf, G. Samsa, A. P. Abernethy, and T. W. LeBlanc, "Integration of electronic patient-reported outcomes into routine cancer care: An analysis of factors affecting data completeness," *JCO Clin. Cancer Informat.*, no. 1, pp. 1–10, Nov. 2017.
[29] C. Liu, A. Talaei-Khoei, D. Zowghi, and J. Daniel, "Data completeness in healthcare: A literature survey," *Pacific Asia J. Assoc. Inf. Syst.*, vol. 9, no. 2, pp. 75–100, 2017.
[30] J. E. DeVoe, R. Gold, P. McIntire, J. Puro, S. Chauvie, and C. A. Gallia, "Electronic health records vs medicaid claims: Completeness of diabetes preventive care data in community health centers," *Ann. Family Med.*, vol. 9, no. 4, pp. 351–358, Jul. 2011.
[31] K. Bhaskaran and L. Smeeth, "What is the difference between missing completely at random and missing at random?" *Int. J. Epidemiol.*, vol. 43, no. 4, pp. 1336–1339, Aug. 2014.
[32] S. Seaman, J. Galati, D. Jackson, and J. Carlin, "Missing at random?" *Stat. Sci.*, vol. 28, no. 2, pp. 257–268, 2013.
[33] S. A. Kulkarni, J. S. Pannu, A. V. Koval, G. J. Merrin, V. Gurupur, A. Nasir, C. King, and T. T. H. Wan, "A brief analysis of key machine learning methods for predicting medicare payments related to physical therapy practices in the United States," *Information*, vol. 12, no. 2, pp. 1–18, 2021.

VARADRAJ P. GURUPUR (Senior Member, IEEE) is currently working as an Associate Professor with the Department of Health Management and Informatics, University of Central Florida. He has more than seven years of teaching experience. He has served as a teacher for two different countries. He has worked in healthcare industry for several years. Based on this work experience and academic training, he is involved in discovering innovative solutions to difficult problems associated with electronic health records. His research interest includes software engineering decision support systems for healthcare and education.

He was a recipient of two international awards, one national award, and several regional and institutional awards.

MUHAMMED SHELLEH (Member, IEEE) received the bachelor's degree in health and biomedical sciences. He is currently a student pursuing his graduate degree in Computer Science with the Department of Computer Science, University of Central Florida.