Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 185 (2021) 177–184

Procedia Computer Science

www.elsevier.com/locate/procedia

Complex Adaptive Systems Conference Theme: Big Data, IoT, and AI for a Smarter Future
Malvern, Pennsylvania, June 16-18, 2021

Machine Learning Models and Big Data Tools for Evaluating Kidney Acceptance

Lirim Ashiku[a,1] Md. Al-Amin[a], Sanjay Madria[b], Cihan Dagli[a]

[a] *Department of Engineering Management and Systems Engineering*
[b] *Department of Computer Science*
Missouri University of Science and Technology, Rolla, MO 65401, USA

Abstract

The rise of on-demand healthcare and the unprecedented growth of electronic health records has given rise to big data opportunities and data analysis using machine learning. The massive and disparate data management using conventional databases is incredibly challenging and expensive to manage. It often requires specialized analytical tools for developing advanced data-driven capabilities and performing data analytics. This paper explores the capability of an open-source framework 'Apache Spark' capable of processing large amounts of data on clusters of nodes to analyze Big data and integrate technologies to provide decision support systems in healthcare settings. Next, we propose machine learning models on top of Apache Spark to expedite the decision-making in allocating organs such as kidney selection for the right candidate, thus increasing donor utilization by locating a recipient within the allotted time. The proposed models help in identifying waitlisted candidates willing to accept kidneys that may otherwise be discarded.

© 2021 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the Complex Adaptive Systems Conference, June 2021.

Keywords: Healthcare; Big data; Machine learning; Apache Spark; Organ procurement.

1877-0509 © 2021 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the Complex Adaptive Systems Conference, June 2021.
10.1016/j.procs.2021.05.019

Nomenclature

EHR	Electronic Health Records
ESRD	End-Stage Renal Disease
KT	Kidney Transplant
TC	Transplant Centre
OPO	Organ Procurement Organization
OPTN	Organ Procurement Transplantation Network
KDPI	Kidney Donor Profile Index
GPU	Graphics Processing Unit
UNOS	United Network for Organ Sharing

1. Introduction

Current technological advancements and artificial intelligence (AI) tools for big data have enabled interactive and collaborative decision making across most industries. Yet, in the healthcare industry, AI integration may still be in the beginning phase. Early phases may be due to the fragmented healthcare system, where high costs are the inherent complexities that may cause misaligned interests and non-collaboration between stakeholders. While time-consuming manual processes and lack of platform interoperability create massive administrative burdens and unnecessary strain in our healthcare system, other industries have steadily eliminated inefficiencies by leveraging AI technologies. Embedding an AI approach to the health industry will contribute to making healthcare more efficient, reduce unnecessary costs, and at the same time, amend clinician's decision-making in optimizing patient care. Though the need to create cost-effective technologies is of the utmost importance, it is associated with extensive resource utilization. Early phases of exploratory data analysis are attributed to the uncertainty of resources looked-for by specialized analytical tools to handle disparate and complex data. Analytical tools may be associated with enormous costs requiring special software to connect to multiple databases or file formats. These tools are needed to divide the data across various sources and use cutting-edge computing power such as graphical processing units (GPUs) to handle faster data processing. However, we could address these limitations with big data methodologies. Big data is used to integrate heterogeneous data generated from various sources, such as electronic health records (EHRs), sensors or actuators, etc., and facilitate accurate and faster medical data analysis in early disease detection [1]. Big data will enable precision medicine, i.e., an emerging health tool aiding transplant surgeons' decision-making to accept organs appropriate for the right candidate [3]. For example, big data analytics will be deployed for massive and disparate deceased-donor kidney historical data to develop AI models for solving a binary class classification problem [22]. Our objective is to quickly identify waitlisted candidates willing to accept kidneys that may otherwise be discarded.

We will train the machine learning model over Apache Spark using prior offer observations for donor-candidate matches. The model will classify offer acceptance, boost acceptance assurance for low transplantation transplant centers (TCs) and serve as a decision aide for marginal kidneys. This should optimize supply chain management and preserve donor kidney degradation caused by both time and transportation factors. The remainder of the paper is organized as follows. An extensive review is described in the second section. Section 3 articulates problem statements, while dataset preprocessing and approach discussion are introduced in the subsequent section. Section 5 and 6 describe algorithm discussion and implementation. Finally, contribution discussion and future work will be presented in the last section.

2. Related Work

The continuous updating of various medical data with high velocity makes it impossible to analyze them with conventional hardware and software platforms [11]. Therefore, several Big data tools such as Apache Mahout, MapReduce, Apache Spark, Apache Flink, and so forth, have been introduced for faster process and extraction of meaningful information, thus enabling big data analytics. In addition to the aforesaid big data platforms' machine learning algorithms, fuzzy inference systems and deep learning methods have been applied in healthcare. These include the diagnosis of diseases, monitoring patient symptoms, tracking chronic diseases, the preventive incidence of contagious diseases, genetic data analytics, personalized medicine, and so on [17]. For instance, a scalable machine learning approach was proposed for Cancer diagnosis using a Hidden Markov Model (HMM) and Gaussian Mixture [12]. A Simultaneously Aided Diagnosis Model (SADM) was presented to aid diagnosis for outpatient care using Support Vector Machine (SVM) and Neural Networks (NN) classifiers [13]. A neuro-fuzzy classifier with recursive feature elimination for top principle components improved classification results for breast cancer prognosis [14]. Moreover, machine learning and deep learning algorithms were also utilized, wherein the former was used for feature extraction, and the latter was used for classification tasks. For example, Convolutional Neural Network (CNN) was employed on ECG signals for extracting discriminative features, and logistic regression was used for cardiac diagnosis classification tasks on the Apache Spark platform [15]. Transfer learning was used to take advantage of pre-trained models in classifying the COVID-19 X-ray images using logistic regression classifier [16]. In the literature, machine learning algorithms have been widely used on big data platforms such as Apache Mahout and Apache Spark. The former has its advantage with distributed storage and scalability. Yet, it provides low processing speed, is unable to process streaming data, and is inefficient with iterative processing.

On the other hand, the latter supports real-time data processing with a swift response time [17]. Apache Spark uses a main memory computing framework based on parallel programming models of Resilient Distributed Datasets (RDDs) and Directed Acyclic Graphs (DAGs). RDDs and DAGs enable data caches to be saved in memory; thus, data processing and training are performed directly from memory. Memory computations allow Spark to perform much faster by avoiding the input-output delay of switching data back and forth from the hard disk. Contrary to Apache Mahout, Apache Spark has been widely used and dominant for big data analytics. For instance, Apache Spark was proposed to detect heart disorders using an electrocardiogram with a Menard algorithm [18]. Alotaibi et al. used the same platform to predict dermal diseases, heart diseases, hypertension, cancer, and diabetes using Naïve Bayes and logistic regression classifiers [19]. A real-time health status prediction system was built on streaming data using a decision tree algorithm [20]. For high performance with low latency, clustering algorithms have also been integrated with the Apache Spark cloud platform for disease diagnosis [11]. Motivated by the literature, this study has proposed the former to facilitate the decision-making about organ procurement.

3. Problem Statement

Nearly 50% of Americans suffer from one or more chronic diseases like end-stage renal disease (ESRD), a kidney functionality degradation where kidneys lose filtering capabilities by exposing one's body to dangerous fluids and accumulated waste in the body [2, 4]. Similarly, about 70% of ESRD patients receive renal replacement therapy through dialysis. However, once they reach irreversible chronic kidney failure, they are listed with local TC as candidates seeking kidney transplants. In comparison to dialysis, kidney transplants offer greater survival and improved quality of life [4-5]. Cox model estimates that even for low-quality kidneys, the transplant's incremental cost-effectiveness is desirable to the alternative [5]. Moreover, the shortage of deceased donor kidney availability results in a long wait time for a transplant, causing high waitlist mortality rates. Long wait times often induce financially competent waitlisted candidates to relocate into the desired geographic area, with TCs having comparatively less waiting times. Lesser wait time may result from higher transplant rate, higher acceptance rates, larger populations leading to larger mortalities, causing an increase in offers to the donor-specific area, etc. [6]. An organ procurement organization (OPO) is a non-profit organization responsible for recovering organs and making electronic offers to TCs. The offers are generated from United Network for Organ Sharing (UNOS) centralized computer network [7]. The computer platform can access all OPOs and TCs and, based on predefined and continuously updated policies, will generate a match-list that matches donors to candidates. Because lower-quality kidney offers may pose a high risk associated with graft failures, TCs may be reluctant to accept. Hence, an organ may go through hundreds of rejections before a TC, if any, will accept. Roughly 59.1% of lesser quality kidneys with a kidney donor profile index (KDPI) greater than 85 are discarded [10]. However, risk-adjusted analysis suggests that candidates are expected to benefit more from such kidneys (KDPI 81-99) as opposed to remaining on dialysis, and of the waitlisted candidates, about 47.8% were willing to accept KDPI > 85 kidneys, [8-10]. Fig. 1

depicts the proposed kidney procurement process that incorporates the AI model to aid transplant surgeons' decision-making.

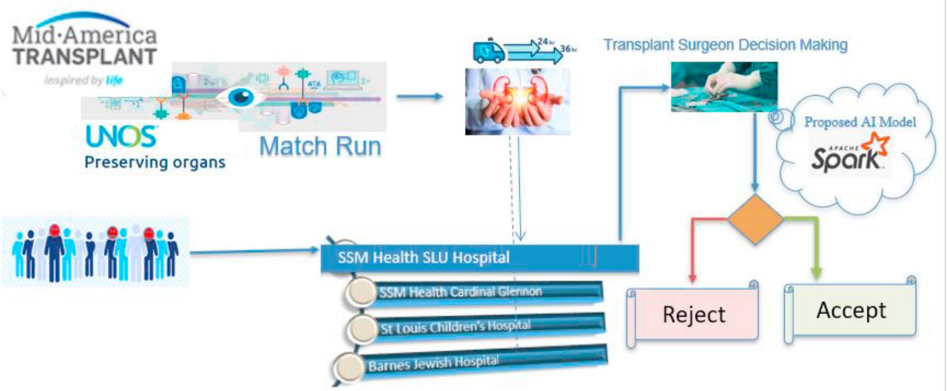


Fig 1. Kidney allocation process and decision making in kidney acceptance.

4. Data Preprocessing

De-identified UNOS match-run datasets used for this study are provided by Abdominal Transplant Center, Saint Louis University School of Medicine, containing both donor, candidates, and many other attributes. The data reflects information about waitlisted candidates for which UNOS found a donor-candidate match. The waitlisted candidates do not represent all waitlisted candidates in the nation, but only the ones for which a kidney offer was made. As indicated earlier, for the waitlisted candidates found as potential transplant recipients by UNOS, there may be numerous rejections and possibly an acceptance. Since observations comprise the donor, candidate, and match-related features, all observations are distinct even though they may repeat candidate features. Historical decisions of accepting or rejecting an offer are used to guide our proposed models. Table 1 depicts the number of observations and features for each of the four datasets. However, some of the datasets are larger than 20 GB, which creates a challenge loading into supported versions of GPUs and pose even a more significant challenge to apply preprocessing techniques. Traditional methods use various chunking techniques to split the dataset into batches allowing smaller batch sizes to be loaded into memory. Yet, this not only prolongs preprocessing but also causes lengthy AI model training. Big data tools offering cluster computing were adopted to overcome memory and speed predicaments, allowing punctual dataset loadings, preprocessing, and model training.

Table I. Datasets and their sizes.

Dataset	Number of Observations	Number of Features
Donor-Deceased	227,733	44
Donor-Disposition	444,231	8
Candidate	912,258	38
Match Run	96,654,094	37

We examined every feature possible distinct value/s, their distribution, and the probable impute value using the scientific registry of transplant recipients (SRTR) data dictionary to deal with missing data. Additionally, we performed multiple imputations with chained equations (MICE). MICE uses algorithms that predict missing values by assigning the feature as a response variable and remaining features as predictors. To amend the uncertain impute values and simultaneously validate imputations, we sought professional guidance from stakeholders at SSM St Louis Hospital to provision applied erudite impute values, thus reducing uncertainty caused by missing values or imputations. Fig. 2 illustrates a pointer to attribute types disclosed in each of the datasets. Features beginning with the 'DON' prefix relate to donor characteristics, whereas the 'PTR' or 'CAN' prefix is associated with patient or candidate characteristics.

After concluding imputation, we attempted predictive modeling as a probabilistic process. Predictive modeling allows forecasting that requires feature analysis and visualizations to facilitate inference about decision-making. With the increased number of features, it becomes computationally costly and challenging to observe relations. After factoring correlated features, we conducted a principal component analysis (PCA) to reduce the number of features yet minimizing information loss. This transformation provided a list of principal components by order of importance, with the first variable preserving the most structure successively followed by variables of lesser principle values. 95% of the variance can be explained by the principal components enumerating 57 variables instead of the original 124 after accounting for the dataset join feature.

We created a binary-class balanced dataset from a combined large dataset following a down-sampling technique. The down-sampled dataset contains 98,000 observations wherein 57 features serve as input and output a response variable to accept or reject an offer.

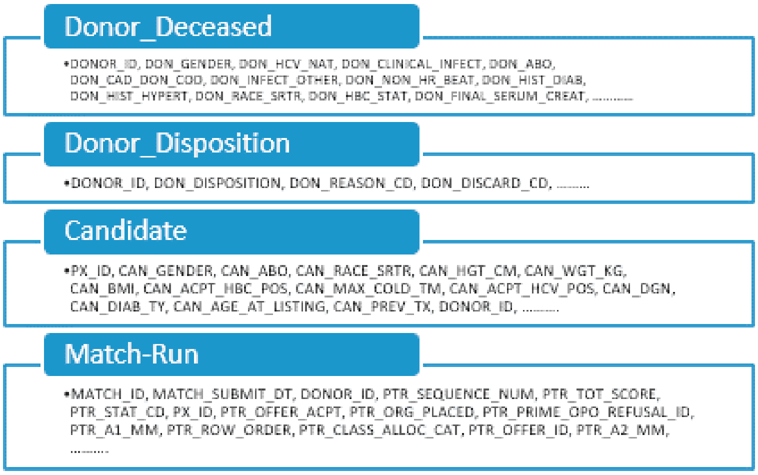

Donor_Deceased
•DONOR_ID, DON_GENDER, DON_HCV_NAT, DON_CLINICAL_INFECT, DON_ABO, DON_CAD_DON_COD, DON_INFECT_OTHER, DON_NON_HR_BEAT, DON_HIST_DIAB, DON_HIST_HYPERT, DON_RACE_SRTR, DON_HBC_STAT, DON_FINAL_SERUM_CREAT, ...
Donor_Disposition
•DONOR_ID, DON_DISPOSITION, DON_REASON_CD, DON_DISCARD_CD, ...
Candidate
•PX_ID, CAN_GENDER, CAN_ABO, CAN_RACE_SRTR, CAN_HGT_CM, CAN_WGT_KG, CAN_BMI, CAN_ACPT_HBC_POS, CAN_MAX_COLD_TM, CAN_ACPT_HCV_POS, CAN_DGN, CAN_DIAB_TY, CAN_AGE_AT_LISTING, CAN_PREV_TX, DONOR_ID, ...
Match-Run
•MATCH_ID, MATCH_SUBMIT_DT, DONOR_ID, PTR_SEQUENCE_NUM, PTR_TOT_SCORE, PTR_STAT_CD, PX_ID, PTR_OFFER_ACPT, PTR_ORG_PLACED, PTR_PRIME_OPO_REFUSAL_ID, PTR_A1_MM, PTR_ROW_ORDER, PTR_CLASS_ALLOC_CAT, PTR_OFFER_ID, PTR_A2_MM, ...

Fig. 2. Attributes associated with each of the datasets.

5. Methodology and Implementation Details

To effectively identify opportunities in the procurement process and increase the number of transplants, we adopted Apache Spark using Python (PySpark) to develop a machine learning model. Apache Spark algorithms run on a Hadoop MapReduce distributed file system. Apache Spark offers various implementations for classification, clustering, and other machine learning algorithms. Using the platform described above, we trained supervised learning algorithms to learn a function that maps input to output based on input-output pairs' observation relations. Once learning occurs, the trained model is then applied to generalize or make predictions on unseen data.

We explored with logistic regression, Naïve Bayes, decision tree, random forest, and multilayer perceptron. The algorithms were trained using both default parameters and tuned hyperparameters like max depth, the minimum number of samples to split the node, a number of features for the best split, impurity function choices, and so forth [21]. Additionally, we used cross-validation for both logistic regression and decision trees. Cross-validations ensure that we test every observation, and in this case, significantly improve performance. The multilayer perceptron classifier developed for the binary classification is based on the feedforward artificial neural network consisting of input, two hidden layers, and the output layer. Nodes in the intermediate layers use a sigmoid activation function, whereas the SoftMax function squashes the output layer.

Similarly, gradient-boosted trees tend to be popular regression methods outperforming deep learning models for tabular data. The gradient-boosted algorithm trains an ensemble of decision trees to minimize the loss function and works well for both binary classification and regression. Moreover, we also explored a decision tree regression algorithm to provide a probability score that generates acceptance for a given donor. The algorithm could help OPOs prioritize TC contact for UNOS match-run candidates to rank the latter based on the probability score.

To enable joined effort for team members, we adopted PySpark API in Google Colab. We installed the open-source implementation 'OpenJDK,' Apache Spark 3.0.1, with Hadoop 2.7 to use distributed processing on a cluster of nodes. We established environment paths to dynamically access Java objects in a Java Virtual Machine using a Python interpreter. We installed the 'Py4J' package to serve as a bridge between Python and Java.

6. Results and Discussions

To substantiate algorithm decision choice for predicting a binary response, classification results for a balanced dataset of 80-20 split is shown in Table II. Initial classification assessments were somewhat similar for most algorithms apart from Logistic Regression sufferings performance, which yielded 56.0% accuracy. However, tuning the hyperparameters and introducing the regularized loss function improved accuracy to 94.89%. This accuracy follows from a 10-fold cross-validation strategy with Logistic Regression. Similarly, a 5-fold cross-validation decision tree model with hyperparameters tuning improved the performance from 90.0% to 93.19%.

Table II. Algorithm assessment.

Algorithm	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.56	0.32	0.56	0.41
Decision Tree	0.90	0.89	0.90	0.90
Random Forest	0.89	0.89	0.88	0.89
Naïve Bayes	0.87	0.87	0.86	0.88
MLP Classifier	0.87	0.87	0.86	0.88
Modified Algorithms				
Logistic Regression Cross-validation	94.89			
Decision Tree Cross-Validation	93.19			
Gradient Boosted Trees	93.96	92.9	94.4	93.89

In this study, a target value of '0' is considered a rejection, and '1' an acceptance. We reviewed confusion matrices for different algorithms to identify the missed opportunities and chose the one with significantly higher false positives. False positives imply that the model predicts acceptance for kidney rejections on test data. Fig. 3 depicts the results obtained from a 5-fold cross-validation decision tree model. The decision tree model yields an accuracy of 93.19% with 741 false positives. Simultaneously, the model predicts a false negative value of 235. This implies that we should have rejected these kidneys. Confusion matrix results led to further analysis of the data to identify whether the kidneys predicted for acceptance were discarded. Further investigation is owing to the selected sample for which a donor kidney may have undergone numerous rejections before being accepted.

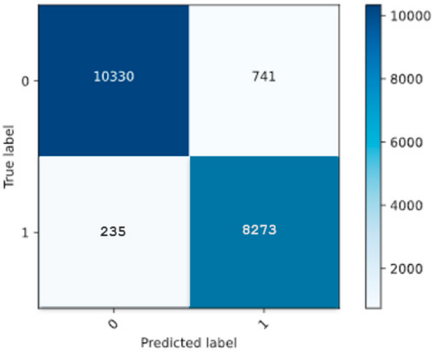
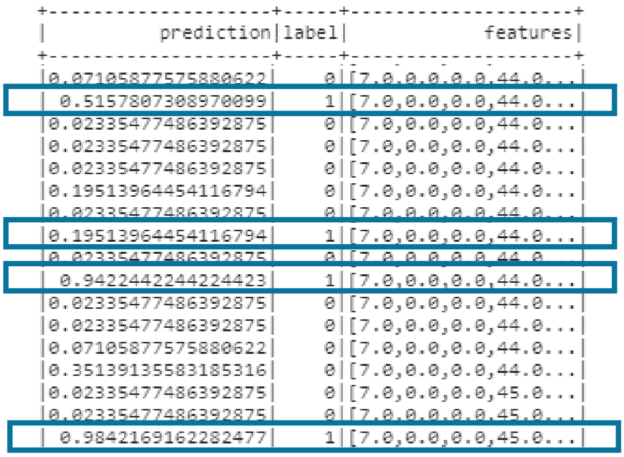


Fig. 3. Confusion matrix for 5-fold cross-validation decision tree model.

A detailed investigation for false positives was conducted using donor disposition attributes to conclude that some of the models predicted acceptance organs were discarded. Although this may not be a true representative of saving donor's kidneys, this would show the genuine model contribution to reducing kidney discard. We may not attribute this with saving a kidney from being discarded because, under rare circumstances, TCs may reject an initially accepted kidney upon arrival for not being as described. In this case, the organ will be made available and offered to local TCs due to the time-sensitive degradation. Nevertheless, the proposed architecture focuses on expediting decision-making, boost acceptance confidence, and reduce kidney discard for this study. Therefore, for rejected and discarded kidneys, identified as false positives by the 5-fold cross-validation model, expedited decision-making would have preserved kidneys from time-sensitive degradations and potentially rescued them from discard. Besides data referencing whether discarded kidneys could have been saved, stakeholders such as transplant surgeons, nephrologists and OPO coordinator will validate the model before it is considered for practice.

A decision tree regression model will determine deceased donor kidney acceptance probability in addition to identifying missed opportunities. This model will serve useful to the OPOs when trying to locate TCs willing to accept available organs. Fig. 4 highlights potential opportunities for the model. This model will systemize OPO to contact TCs based on UNOS match ranking and the likelihood of acceptance shown in the 'prediction' column in Fig. 5. Moreover, OPO staff can specify a threshold value for prediction and begin promoting organs to TCs based on the defined value.



```
+-------------------+-----+--------------------+
|         prediction|label|            features|
+-------------------+-----+--------------------+
|0.07105877575880622|    0|[7.0,0.0,0.0,44.0...|
| 0.5157807308970099|    1|[7.0,0.0,0.0,44.0...|
|0.02335477486392875|    0|[7.0,0.0,0.0,44.0...|
|0.02335477486392875|    0|[7.0,0.0,0.0,44.0...|
|0.02335477486392875|    0|[7.0,0.0,0.0,44.0...|
|0.19513964454116794|    0|[7.0,0.0,0.0,44.0...|
|0.02335477486392875|    0|[7.0,0.0,0.0,44.0...|
|0.19513964454116794|    1|[7.0,0.0,0.0,44.0...|
|0.02335477486392875|    0|[7.0,0.0,0.0,44.0...|
| 0.9422442244224423|    1|[7.0,0.0,0.0,44.0...|
|0.02335477486392875|    0|[7.0,0.0,0.0,44.0...|
|0.02335477486392875|    0|[7.0,0.0,0.0,44.0...|
|0.07105877575880622|    0|[7.0,0.0,0.0,44.0...|
|0.35139135583185316|    0|[7.0,0.0,0.0,44.0...|
|0.02335477486392875|    0|[7.0,0.0,0.0,45.0...|
|0.02335477486392875|    0|[7.0,0.0,0.0,45.0...|
| 0.9842169162282477|    1|[7.0,0.0,0.0,45.0...|
```

Fig. 4. Results with decision tree regression model.

7. Conclusions and Future Work

In this paper, big data tools have been analyzed and used to develop machine learning models to aid transplant surgeons' decision-making regarding kidney acceptance. The use of big data tools enabled handling big datasets that was not possible using conventional data analytics platforms and operated at significantly improved speed. 'Apache Spark' was adopted to analyze a large-scale disparate dataset and created predictive machine learning models. The classification models will facilitate expedited decision–making in a nearly real-time manner for transplant surgeons. Given that we wanted to identify missed opportunities, we didn't see it necessary to plot the area under the receiver operating characteristics (AUROC). By investigating false positives, the model may have uncovered missed opportunities for rejected and discarded organs that may have been otherwise accepted. Furthermore, a regression model will aid in systemizing the contact to TCs in promoting the organ by providing the likelihood in addition to the acceptance decision. Provided that the models were trained on dataset observations for which UNOS found donor-candidate match, we propose the models be used to complement current kidney allocation process and not as a replacement of the existing allocation process. Future work involves developing predictive models through use of GPUs and big data-parallel computing framework and perform close comparison in terms of both speed and accuracy. Additionally, we can implement ensembled methods to combine the results from different algorithms to yield a final prediction and compare the constituent classifiers' accuracy.

Acknowledgments

This work was supported by the Missouri University of Science and Technology and Saint Louis University. Any opinions, findings, and conclusions or recommendations expressed in this paper are attributed to the authors and do not reflect any organization's views. The authors are grateful to the Abdominal Transplant Center, Saint Louis University School of Medicine, for providing the dataset. They also thank Dr. Krista Lentine and Dr. Mark Schnitzler for their insightful comments and suggestions in this research.

References

[1] Chen, Min, Yixue Hao, Kai Hwang, Lu Wang, and Lin Wang. (2017) "Disease prediction by machine learning over big data from healthcare communities," *IEEE Access* **5**: 8869-8879.

[2] Groves, Peter, Basel Kayyali, David Knott, and Steve Van Kuiken. (2016) "The big data revolution in healthcare: Accelerating value and innovation."

[3] Zhang, Yin, Meikang Qiu, Chun-Wei Tsai, Mohammad Mehedi Hassan, and Atif Alamri. (2015) "Health-CPS: Healthcare cyber-physical system assisted by cloud and big data." *IEEE Systems Journal*, 11, no. **1**: 88-95.

[4] Davis, Ashley E., Sanjay Mehrotra, Lisa M. McElroy, John J. Friedewald, Anton I. Skaro, Brittany Lapin, Raymond Kang, Jane L. Holl, Michael M. Abecassis, and Daniela P. Ladner. (2014) "The extent and predictors of waiting time geographic disparity in kidney transplantation in the United States." *Transplantation,* 97, no. 10: 1049-1057.

[5] Katzman, Jared L., Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. (2018) "DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network." *BMC Medical Research Methodology* **18**, no. 1: 24.

[6] Merion, Robert M., Mary K. Guidinger, John M. Newmann, Mary D. Ellison, Friedrich K. Port, and Robert A. Wolfe. (2004) "Prevalence and outcomes of multiple-listing for cadaveric kidney and liver transplantation." *American Journal of Transplantation* **4**, no. 1: 94-100.

[7] Organ procurement organizations: Increasing organ donations. (2020). Retrieved September 14, 2020, from https://unos.org/transplant/opos-increasing-organ-donation/

[8] Karnofsky Performance Status Scale. (n.d.). Retrieved September 14, 2020, from https://www.mdcalc.com/karnofsky-performance-status-scale

[9] Bui, Kevin, Vikram Kilambi, and Sanjay Mehrotra. (2019) "Functional status-based risk–benefit analyses of high-KDPI kidney transplant versus dialysis." *Transplant International,* **32**, no. 12: 1297-1312.

[10] Hart, A., J. M. Smith, M. A. Skeans, S. K. Gustafson, D. E. Stewart, W. S. Cherikh, J. L. Wainright et al. (2017) "OPTN/SRTR 2015 annual data report: kidney." *American Journal of Transplantation*, **17**: 21-116.

[11] Chen J, Li K, Rong H, Bilal K, Yang N, Li K. (2018) "A disease diagnosis and treatment recommendation system based on big data mining and cloud computing." *Information Sciences*, 2018 Apr 1; 435: 124-49.

[12] Manogaran G, Vijayakumar V, Varatharajan R, Kumar PM, Sundarasekar R, and Hsu CH. (2018) "Machine learning based big data processing framework for cancer diagnosis using hidden Markov model and GM clustering. Wireless personal communications." 2018 Oct 1;**102 (3)**: 2099-116.

[13] Hu Y, Duan K, Zhang Y, Hossain MS, Rahman SM, and Alelaiwi A. (2018) "Simultaneously aided diagnosis model for outpatient departments via healthcare big data analytics." *Multimedia Tools and Applications*, 2018 Feb 1; **77 (3)**: 3729-43.

[14] Azar AT, and Hassanien AE. (2015) "Dimensionality reduction of medical big data using neural-fuzzy classifier." *Soft Computing*, 2015 Apr 1; **19 (4)**: 1115-27.

[15] Tun ZM, and Khine MA. (2020) "Cardiac Diagnosis Classification Using Deep Learning Pipeline on Apache Spark." *17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology,* (ECTI-CON), 2020 Jun 24, pp. 743-746, IEEE.

[16] Benbrahim H, Hachimi H, and Amine A. (2020) "Deep transfer learning with apache spark to detect covid-19 in chest x-ray images." *Romanian Journal of Information Science and Technology*, 2020 Jan 1; **23**: S117-29.

[17] Nazari E, Shahriari MH, and Tabesh H. (2019) "BigData Analysis in Healthcare: Apache Hadoop, Apache spark and Apache Flink." *Frontiers in Health Informatics*, 2019 Jul 27; **8 (1)**: 14.

[18] Carnevale L, Celesti A, Fazio M, Bramanti P, and Villari M. (2017) "Heart disorder detection with menard algorithm on apache spark." *European Conference on Service-Oriented and Cloud Computing*, 2017 Sep 27, 229-237, Springer, Cham.

[19] Alotaibi S, Mehmood R, Katib I, Rana O, and Albeshri A. Sehaa. (2020) "A big data analytics tool for healthcare symptoms and diseases detection using Twitter, Apache Spark, and Machine Learning." *Applied Sciences*, 2020 Jan; **10 (4)**: 1398.

[20] Nair LR, Shetty SD, and Shetty SD. (2018) "Applying spark based machine learning model on streaming big data for health status prediction." *Computers & Electrical Engineering*, 2018 Jan 1; **65**: 393-9.

[21] Bickerton, C. A. (2018) "Beginner's guide to decision tree classification." (2018, August 1). Retrieved November 3, 2020, from https://towardsdatascience.com/a-beginners-guide-to-decision-tree-classification-6d3209353ea

[22] Ristevski, Blagoj, and Ming, Chen. (2018) "Big data analytics in medicine and healthcare." *Journal of Integrative Bioinformatics* 15, no. 3.