Prediciendo EICH(Enfermedad de injerto contra huésped) en niños después del trasplante de medula osea con técnicas de Machine Learning

Lázaro Gibert García

Introducción

En la actualidad los métodos de **Machine Learning** han demostrado ser de gran utilidad en varios campos de nuestra vida diaria. En el campo de la medicina estos métodos aún sufren de escepticismo debido a la falta de explicabilidad de los que sufren algunos, ya que los especialistas médicos no son capaces de comprender el porqué de las decisiones que toma el método. Aún así se ha logrado una importante aceptación de estos algoritmos con la implementación de modelos que aportan interpretabilidad [8, 18] y se han desarrollado varios trabajos en el campo de la oncología.

La enfermedad de injerto contra huésped(EICH, o GVHD por la siglas en inglés de Graft-Versus-Host-Disease) es una complicación médica común en determinados alotrasplantes de tejidos. Se asocia principalmente al trasplante de células pluripotenciales hematopoyéticas, comúnmente denominado trasplante de médula ósea. En última instancia esta complicación es debida a que las células inmunes presentes en el tejido trasplantado reconocen al receptor del trasplante(el hospedador) como extraño. Las células inmunes trasplantadas una vez activadas atacan a las células del receptor causando la enfermedad. En algunas ocasiones puede presentarse la enfermedad de injerto contra huésped después de una transfusión sanguínea. Esta enfermedad es potencialmente mortal y solo ocurre si el receptor no recibe sus propias células, es decir, si las recibe de un donante.

El objetivo de este trabajo es aplicar una serie de modelos de **Machine Learning** a un dataset disponible públicamente que contiene datos sobre el trasplante de médula ósea en niños y así lograr predecir si el paciente desarrollará EICH o no.

Antecedentes

El desarrollo de técnicas de Machine Learning para apoyar a los especialistas oncológicos en su labor es algo que ha llamado la atención de varios especialistas en esta área de la inteligencia artificial, esto se puede apreciar en [15, 19, 16] donde se generan una estrategia de inducción de reglas para predecir si un paciente sobrevivirá al trasplante de médula ósea y los factores que hacen posibles estos resultados. No sólo se han utilizado métodos de inducción de reglas para estudiar esta área también se han utilizado métodos basados en árboles de decisión como se puede ver en [9] donde de genera un algoritmo que se puede aplicar en el análisis de supervivencia de los pacientes. El área de interés no se limita a la predicción de supervivencia de los pacientes, existen estudios para predecir complicaciones específicas al trasplante de médula ósea como lo es la Enfermedad de Injerto Contra Huésped(EICH), una enfermedad muy común y potencialmente mortal [1, 10, 3, 12]. Para el desarrollo los estudios y predicción de la EICH se han utilizado varias fuentes de datos, desde Registros electrónicos de salud [17] hasta resultados de laboratorio almacenados en forma tabular [7]. Además se ha trabajado en formas más específicas de este enfermedad como en [14] donde se predice la EICH severo y en [5] se predice la EICH crónica.

Descripción del dataset

El conjunto de datos describe pacientes pediátricos con varias enfermedades hematológicas: trastornos malignos(p. ej., leucemia linfoblástica aguda, leucemia mielógena aguda, leucemia mielógena crónica, síndrome mielodisplásico)

y casos no malignos(p. ej., anemia aplásica grave, anemia de Fanconi, con adrenoleucodistrofia ligada al cromosoma X). Todos los pacientes se sometieron al trasplante de células madre hematopoyéticas de un donante alogénico no emparentado. Este dataset se puede encontrar publicado en https://www.kaggle.com/datasets/adamgudys/bone-marrow-transplant-children?resource=download. El conjunto contiene 187 ejemplos caracterizados por 37 atributos. El significado de las características se puede apreciar en el cuadro. 1

Nombre de la variable	**Descripción**	**Tipo de dato**
donor_age	Edad del donante en el momento de la aféresis de células madre hematopoyéticas.	Numérico
donoragebelow_35	La edad del donante es inferior a 35 años.	Nominal(sí, no)
donor_ABO	Grupo sanguíneo ABO del donante de células madre hematopoyéticas.	Nominal(0, A, B, AB)
donor_CMV	Presencia de infección por citomegalovirus en el donante de células madre hematopoyéticas antes del trasplante.	Nominal(presente, ausente)
recipient_age	Edad del receptor de células madre hematopoyéticas en el momento del trasplante.	Numérico
recipientagebelow_10	¿La edad del destinatario es inferior a 10?	Nominal(sí, no)
recipientageint	Edad del destinatario discretizada a intervalos.	Nominal(0,5], (5, 10], (10, 20]
recipient_gender	Sexo del destinatario.	Nominal(femenino, masculino)
recipientbodymass	Masa corporal del receptor de células madre hematopoyéticas en el momento del trasplante.	Numérico
recipient_ABO	Grupo sanguíneo ABO del receptor de células madre hematopoyéticas.	Nominal(0, A, B, AB)
recipient_rh	Presencia del factor Rh en los glóbulos rojos del receptor.	Nominal(más, menos)
recipient_CMV	Presencia de infección por citomegalovirus en el donante de células madre hematopoyéticas antes del trasplante.	Nominal(presente, ausente)
disease	Tipo de enfermedad.	Nominal(LLA, AML, crónica, no maligna, linfoma)
disease_group	Tipo de enfermedad.	Nominal(maligna, no maligna)
gender_match	Compatibilidad del donante y el receptor según su género.	Nominal(femenino a masculino, otro)
ABO_match	Compatibilidad del donante y el receptor de células madre hematopoyéticas según el grupo sanguíneo ABO.	Nominal(emparejado, no coincidente)
CMV_status	Compatibilidad serológica del donante y el receptor de células madre hematopoyéticas según la infección por citomegalovirus antes del trasplante.	Numérico(a mayor valor, menor compatibilidad)
HLA_match	Compatibilidad de antígenos del principal complejo de histocompatibilidad del donante y el receptor de células madre hematopoyéticas.	Nominal(10/10, 9/10, 8/10, 7/10)
HLA_mismatch	HLA coincidente o no coincidente	Nominal
antigen	En cuántos antígenos hay una diferencia entre el donante y el receptor.	Numérico(0-3)
allel	En cuántos alelos hay una diferencia entre el donante y el receptor.	Numérico(0-4)

Nombre de la variable	Descripción	Tipo de dato
HLAgroup1	El tipo de diferencia entre el donante y el receptor.	Nominal(coincidencia de HLA, un antígeno, un alelo, célula DRB1, dos alelos o alelo+antígeno, dos antígenos+alelo, no coincidentes)
risk_group	Grupo de riesgo.	Nominal(alto, bajo)
stemcellsource	Fuente de células madre hematopoyéticas.	Nominal(sangre periférica, médula ósea)
txpostrelapse	¿El segundo trasplante de médula ósea después de la recaída?	Nominal(sí, no)
CD34x1e6per_kg	CD34kgx10d6 - Dosis de células CD34+ por kg de peso corporal del receptor (10ˆ6/kg).	Numérico
CD3x1e8per_kg	Dosis de células CD3+ por kg de peso corporal del receptor (10ˆ8/kg).	Numérico
CD3toCD34_ratio	Proporción de células CD3+ a células CD34+.	Numérico
ANC_recovery	Tiempo de recuperación de neutrófilos definido como recuento de neutrófilos >0,5 x 10ˆ9/L.	Numérico
PLT_recovery	Tiempo de recuperación de plaquetas definido como recuento de plaquetas >50000/mm3.	Numérico
acuteGvHDIIIIIIV	Desarrollo de enfermedad aguda de injerto contra huésped en estadio II, III o IV.	Nominal(sí, no)
acuteGvHDIII_IV	Desarrollo de enfermedad aguda de injerto contra huésped en estadio III o IV.	Nominal(sí, no)
timetoacuteGvHDIII_IV	Tiempo hasta el desarrollo de la enfermedad de injerto contra huésped aguda en estadio III o IV	Numérico
extensivechronicGvHD	Desarrollo de enfermedad crónica extensa de injerto contra huésped.	Nominal(sí, no)
relapse	Recaída de la enfermedad.	Nominal(sí, no)
survival_time	Tiempo de observación (si está vivo) o tiempo hasta el evento (si está muerto) en días	Numérico
survival_status	Estado de supervivencia.	Numérico(0 - vivo, 1 - muerto)

Cuadro 1: Descripción del dataset

Modelos

La motivación de este trabajo es apoyar a los especialistas brindando una herramienta que prediga si un paciente con ciertas características podría desarrollar EICH, por lo que este constituye un problema de clasificación de clases y por lo tanto los modelos a aplicar son aquellos que se especializan en la solución de este tipo de problemas.

XGBoost

El clasificador XGBoost es un algoritmo de Machine Learning que se aplica a datos estructurados y tabulares. Es una implementación de árboles de decisión potenciados por gradientes diseñados para la velocidad y el rendimiento, además es un algoritmo de aumento de gradiente extremo [4]. Y eso significa que es un gran algoritmo de aprendizaje automático con muchas partes. XGBoost funciona con conjuntos de datos grandes y complicados y a su vez es una

técnica de modelado de conjunto.

Este modelo es un método de aprendizaje conjunto(ensemble learning method). A veces, puede que no sea suficiente confiar en los resultados de un solo modelo de aprendizaje automático. El aprendizaje en conjunto ofrece una solución sistemática para combinar el poder predictivo de múltiples learners. El resultado es un solo modelo que da la salida agregada de varios modelos.

Los modelos que forman el conjunto, también conocidos como learners base, pueden ser del mismo algoritmo de aprendizaje o de diferentes algoritmos de aprendizaje. El embolsado, el impulso, la generalización de la pila y las mezclas expertas son los modelos de aprendizaje de conjunto más utilizados. Sin embargo, embolsar y aumentar son dos conjuntos de aprendizaje muy elogiados. Aunque estas dos técnicas se pueden usar con varios modelos estadísticos, el uso más predominante ha sido con árboles de decisión.

Random forest

Random forest(o random forests) también conocidos en castellano como Bosques Aleatorios es una combinación de árboles predictores tal que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos [2]. Es una modificación sustancial de bagging que construye una larga colección de árboles no correlacionados y luego los promedia.

El algoritmo para inducir un random forest fue desarrollado por Leo Breiman y Adele Cutler y Random forests es su marca de fábrica. El término aparece de la primera propuesta de Random decision forests, hecha por Tin Kam Ho de Bell Labs en 1995. El método combina la idea de bagging de Breiman y la selección aleatoria de atributos, introducida independientemente por Ho, Amit y Geman, para construir una colección de árboles de decisión con variación controlada. La selección de un subconjunto aleatorio de atributos es un ejemplo del método random subspace, el que, según la formulación de Ho, es una manera de llevar a cabo la discriminación estocástica propuesta por Eugenio Kleinberg.

En muchos problemas el rendimiento del algoritmo random forest es muy similar a la del boosting, y es más simple de entrenar y ajustar. Como consecuencia, el Random forest es popular y ampliamente utilizado.

Regresión Logística

La regresión logística es una técnica de clasificación prestada por el aprendizaje automático del campo de la estadística. La regresión logística es un método estadístico para analizar un conjunto de datos en el que hay una o más variables independientes que determinan un resultado. La intención detrás del uso de la regresión logística es encontrar el modelo que mejor se ajuste para describir la relación entre la variable dependiente y la independiente [6].

La regresión logística es una técnica de clasificación utilizada en el aprendizaje automático. Utiliza una función logística para modelar la variable dependiente. La variable dependiente es de naturaleza dicotómica, es decir, solo podría haber dos clases posibles(p. ej.: si el cáncer es maligno o no). Como resultado, esta técnica se utiliza al tratar con datos binarios.

Gradient boosting

Gradient boosting o Potenciación del gradiente, es una técnica de aprendizaje automático utilizado para el análisis de la regresión y para problemas de clasificación estadística, el cual produce un modelo predictivo en forma de un conjunto de modelos de predicción débiles, típicamente árboles de decisión. Construye el modelo de forma escalonada como lo hacen otros métodos de boosting, y los generaliza permitiendo la optimización arbitraria de una función de pérdida diferenciable [11].

La idea de la potenciación del gradiente fue originada en la observación realizada por Leo Breiman en donde el Boosting puede ser interpretado como un algoritmo de optimización en una función de coste adecuada. Posteriormente Jerome H. Friedman desarrolló algoritmos de aumento de gradiente de regresión explícita, simultáneamente con la perspectiva más general de potenciación del gradiente funcional de Llew Mason, Jonathan Baxter, Peter Bartlett y Marcus Frean. En sus últimos dos trabajos presentaron la visión abstracta de los algoritmos de potenciación como algoritmos iterativos de descenso de gradientes funcionales. Es decir, algoritmos que optimizan una función de coste

sobre el espacio de función mediante la elección iterativa de una función(hipótesis débil) que apunta en la dirección del gradiente negativo. Esta visión de gradiente funcional de potenciación ha llevado al desarrollo de algoritmos de potenciación en muchas áreas del aprendizaje automático y estadísticas más allá de la regresión y la clasificación.

AdaBoost

Ada-boost o Adaptive Boosting es uno de los clasificadores potenciadores de conjuntos propuestos por Yoav Freund y Robert Schapire en 1996. Combina múltiples clasificadores para aumentar la precisión de los clasificadores. AdaBoost es un método de conjunto iterativo. El clasificador AdaBoost crea un clasificador fuerte al combinar varios clasificadores de bajo rendimiento para que obtenga un clasificador fuerte de alta precisión. El concepto básico detrás de Adaboost es establecer los pesos de los clasificadores y entrenar la muestra de datos en cada iteración para garantizar predicciones precisas de observaciones inusuales. Cualquier algoritmo de aprendizaje automático puede usarse como clasificador base si acepta pesos en el conjunto de entrenamiento. AdaBoost es fácil de implementar. Corrige iterativamente los errores del clasificador débil y mejora la precisión al combinar aprendices débiles. Es posible usar muchos clasificadores básicos con AdaBoost y no es propenso al sobreajuste [13]. Es sensible a los datos de ruido. Se ve muy afectado por los valores atípicos porque intenta encajar perfectamente en cada punto y es más lento en comparación con XGBoost.

Resultados

Para lograr predecir la EICH se utilizaron varios modelos de clasificación y se escogió el mejor modelo prestando atención a las métricas que se utilizan en este tipo de problemas(accuracy, precision, recall, f1-score y matriz de confusión). También se realizó la validación de cada modelo utilizando la validación cruzada con el método RepeatedKFold con parámetros nsplits=10 y nrepeats=5 y la optimización de hiperparámetros buscando encontrar el modelo que mejor se comportara en el conjunto de prueba utilizando como métrica el **puntaje f1(f1 score)** que es un promedio ponderado de precisión y sensibilidad(recall) y luego se midió el error en el conjunto test con la métrica **fbeta** que es un f1 score con la particularidad de que si el β es mayor que 1 le da mayor peso a la presición y en caso de que β sea menor a 1 presta mayor atención al recall, en este caso es más importante la presición con la cual los modelos aciertan. En la figura 1 se muestra el comportamiento de los errores de los modelos en el conjunto de prueba. El dataset tiene celdas con valores en blanco por lo que para implementar los modelos se eliminaron esas filas, eliminando un total de 45 filas y disminuyendo el número de observaciones del dataset de 187 a 142, considerando que en el campo de la medicina es más apropiado eliminar estos datos que introducir nuevos no reales.

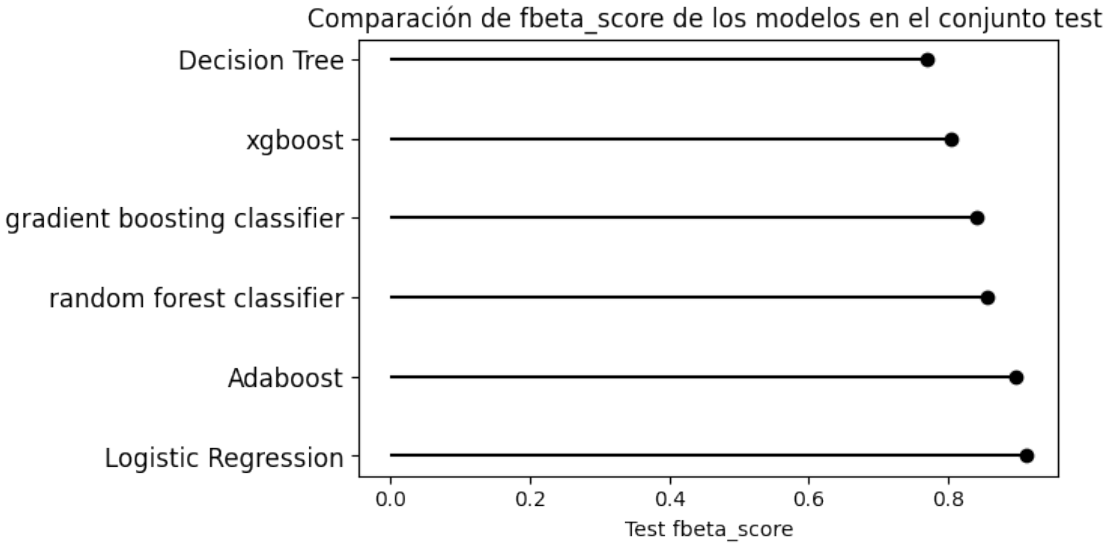

Figura 1: Comparación entre los puntajes f1 en el conjunto de prueba

Métricas	Decision Tree	Random forest	Xgboost	Regresion Logistica	Gradient boosting	Adaboost
Accuracy	0.68	0.72	0.72	0.83	0.72	0.79
f1score	0.64	0.66	0.63	0.66	0.65	0.68
fbeta	0.77	0.85	0.81	0.91	0.84	0.90

Cuadro 2: Métricas de los modelos

En el cuadro 2 se puede apreciar las métricas más representativas de todos los modelos analizados y se puede notar que los de mejor comportamiento son la Regresón Logística, el Adaboost y el Random Forest, pero la Regresón Logística no solo tiene un mayor accuracy, además cuenta con el mayor fbeta en el momemnto de medir el error en el conjunto test por lo que termina siendo el mejor modelo. Como resultado final se obtiene un modelo capaz de predecir el futuro desarrollo de la EICH en pacientes después de ser sometido al procedimiento de trasplante de médula ósea con un accuracy del 83 % y un fbeta del 91 %.

Conclusiones

La predicción de la aparición de la Enfermedad de Injerto Contra Huésped constituye una herramienta de gran apoyo para los especialistas oncológicos ya que les permite tomar las medidas necesarias para evitar que se desarrolle o para tratarla lo mejor posible. Tras haber aplicado varios modelos de clasificación en el dataset, el modelo de Regresión Logística resultó ser el más apropiado para esta tarea, resultando tener un accuracy alto y además de ser el que mejor se comporta cuando se mide el error en el conjunto test ya que al tratar con la salud de los pacientes de vital importancia que el modelo acierte la mayor cantidad de veces posibles en su predicción.

Referencias

[1] Yasuyuki Arai, Tadakazu Kondo, Kyoko Fuse, Yasuhiko Shibasaki, Masayoshi Masuko, Junichi Sugita, Takanori Teshima, Naoyuki Uchida, Takahiro Fukuda, Kazuhiko Kakihana, Yukiyasu Ozawa, Tetsuya Eto, Masatsugu Tanaka, Kazuhiro Ikegame, Takehiko Mori, Koji Iwato, Tatsuo Ichinohe, Yoshinobu Kanda, and Yoshiko Atsuta. Using a machine learning algorithm to predict acute graft-versus-host disease following allogeneic transplantation. *Blood Advances*, 3:3626–3634, 2019.

[2] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[3] Giovanni Caocci, Roberto Baccoli, Adriana Vacca, Angela Mastronuzzi, Alice Bertaina, Eugenia Piras, Roberto Littera, Franco Locatelli, Carlo Carcassi, and Giorgio La Nasa. Comparison between an artificial neural network and logisticregression in predicting acute graft-vs-host disease after unrelated donor hematopoietic stem cell transplantation in thalassemia patients. *Experimental Hematology*, 38:426–433, 2010.

[4] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[5] Jocelyn S. Gandelman, Michael T. Byrne, Akshitkumar M. Mistry, Hannah G. Polikowsky, Kirsten E. Diggins, Heidi Chen, Stephanie J. Lee, Mukta Arora, Corey Cutler, Mary Flowers, Joseph Pidala, Jonathan M.Irish, and Madan H. Jagasia. Machine learning reveals chronic graft-versus- host disease phenotypes and stratifies survival after stem cell transplant for hematologic malignancies. *Haematologica*, 104:189–196, 2019.

[6] Joseph M. Hilbe. *Text in Statical Science Logistic Regression Models.* Taylor and Francis Group, ISBN 978-042-914-913-9, 2009.

[7] Makoto Iwasaki, Junya Kanda, Yasuyuki Arai, Takayuki Ishikawa Tadakazu Kondo, Yasunori Ueda, Kazunori Imada, Takashi Akasaka, Akihito Yonezawa, Kazuhiro Yago, Masaharu Nohgawa, Naoyuki Anzai, Toshinori Moriguchi, Toshiyuki Kitano, Mitsuru Itoh, Nobuyoshi Arima, Tomoharu Takeoka, Mitsumasa Watanabe, Hirokazu Hirata, Kosuke Asagoe, Isao Miyatsuka, Le My An, Masanori Miyanishi, and Akifumi Takaori-Kondo. Establishment of a predictive model for gvhd-free, relapse-free survival after allogeneic hsct using ensemble learning. *Blood Advances*, 6:2618–2627, 2022.

[8] Maxim S. Kovalev, Lev V. Utkin, and Ernest M. Kasimov. Survlime: A method for explaining machine learning survival models. *Knowledge-Based Systems*, 203:1–20, 2020.

[9] Malgorzata Kretowska. Tree-based models for survival data with competing risks. *Computer Methods and Programs in Biomedicine*, 159:185–198, 2018.

[10] Xueou Liu, Yigeng Cao, Ye Guo, Xiaowen Gong, Yahui Feng, Yao Wang, Mingyang Wang, Mengxuan Cui, Wenwen Guo, Luyang Zhang, Ningning Zhao, Xiaoqiang Song, Xuetong Zheng, Xia Chen, Qiujin Shen, Song Zhang, Zhen Song, Linfeng Li, Sizhou Feng, Mingzhe Han, Xiaofan Zhu, Erlie Jiang, and Junren Chen. Dynamic forecasting of severe acute graft-versus-host disease after transplantation. *Natural computational science*, 2:153–159, 2022.

[11] Alexey Natekin and Alois Knoll. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7:1–21, 2013.

[12] Cirruse Salehnasab, Abbas Hajifathali, Farkhondeh Asadi, Sayeh Parkhideh, Alireza Kazemi, Arash Roshanpoor, Mah shid Mehdizadeh, Maria Tavakoli-Ardakani, and Elham Roshandel. An intelligent clinical decision support system for predicting acute graft-versus-host disease (agvhd) following allogeneic hematopoietic stem cell transplantation. *Journal of Biomedical Physics and Engineering*, 11:345–356, 2021.

[13] Robert E. Schapire. Explaining adaboost. *Empirical Inference*, pages 37–52, 2013.

[14] Meng-Zhu Shen, Shen-Da Hong, Rui Lou, Rui-Ze Chen, Xiao-Hui Zhang, Lan-Ping Xu, Yu Wang, Chen-Hua Yan, Huan Chen, Yu-Hong Chen, Wei Han, Feng-Rong Wang, Jing-Zhi Wang, Kai-Yan Liu, Xiao-Jun Huang, and Xiao-Dong Mo. A comprehensive model to predict severe acute graft-versus-host disease in acute leukemia patients after haploidentical hematopoietic stem cell transplantation. *Experimental Hematology and Oncology*, 11:1–10, 2022.

[15] Marek Sikora, Łukasz Wróbel, and Adam Gudys. Guider: A guided separate-and-conquer rule learning in classification, regression, and survival settings. *Knowledge-Based Systems*, 173:1–14, 2019.

[16] Marek Sikora, Łukasz Wróbel, Monika Mielcarek, and Krzysztof Kalwak. Application of rule induction to discover survival factors of patients after bone marrow transplantation. *JOURNAL OF MEDICAL INFORMATICS and TECHNOLOGIES*, 22:35–53, 2013.

[17] Shengpu Tang, Grant T. Chappell, Amanda Mazzoli, Muneesh Tewari, Sung Won Choi, and Jenna Wiens. Predicting acute graft-versus-host disease using machine learning and longitudinal vital sign data from electronic health records. *JCO Clinical Cancer Informatics*, 4:128–135, 2020.

[18] Lev V. Utkin, Egor D. Satyukov, and Andrei V. Konstantinov. Survnam: The machine learning survival model explanation. *Neural Networks*, 147:81–102, 2022.

[19] Łukasz Wróbel, Adam Gudys, and Marek Sikora. Learning rule sets from survival data. *BMC Bioinformatics*, 18(285):1–13, 2017.