UNIVERSIDADE ESTADUAL DE MONTES CLAROS

Laércio Ives Santos

Adaptação de algoritmos híbridos baseados em Aprendizagem de Máquinas para aplicação em problemas na área de Saúde com bases de dados desbalanceadas

Montes Claros
2021

Laércio Ives Santos

Adaptação de algoritmos híbridos baseados em Aprendizagem de Máquinas para aplicação em problemas na área de Saúde com bases de dados desbalanceadas

Tese apresentada ao Programa de Pós-graduação em Ciências em Saúde da Universidade Estadual de Montes Claros-Unimontes, como parte das exigências para obtenção do título de Doutor em Ciências da Saúde.

Área de Concentração: Saúde Coletiva
Orientador: Prof. Dr. Marcos Flávio Silveira Vasconcelos D'Angelo
Coorientador: Prof. Dr. João Batista Mendes

Montes Claros
2021

S237a

Santos, Laércio Ives.

Adaptação de algoritmos híbridos baseados em Aprendizagem de Máquinas para aplicação em problemas na área de saúde com bases de dados desbalanceadas [manuscrito] / Laércio Ives Santos. – Montes Claros, 2021.

57 f. : il.

Inclui Bibliografia.

Tese (Doutorado) - Universidade Estadual de Montes Claros - Unimontes, Programa de Pós-Graduação em Ciências da Saúde /PPGCS, 2021.

Orientador: Prof. Dr. Marcos Flávio Silveira Vasconcelos D'Angelo.

Coorientador. Prof. Dr. João Batista Mendes.

1. Saúde pública. 2. Aprendizado de Máquina – Inteligência Artificial. 3. Prevenção de quedas em idosos. 4. Acidente vascular cerebral. I. D'Angelo, Marcos Flávio Silveira Vasconcelos. II. Mendes, João Batista. III. Universidade Estadual de Montes Claros. IV. Título.

Catalogação Biblioteca Central Professor Antônio Jorge

UNIVERSIDADE ESTADUAL DE MONTES CLAROS-UNIMONTES

Reitor: Prof. Dr. Antônio Avilmar Souza

Vice-reitora: Profa. Dra. Ilva Ruas de Abreu

Pró-reitora de Pesquisa: Profa. Dra. Clarice Diniz Alvarenga Corsato

Coordenadoria de Acompanhamento de Projetos: Prof. Dr. Virgílio Mesquita Gomes

Coordenadoria de Iniciação Científica: Profa. Dra. Maria Alice Ferreira dos Santos

Coordenadoria de Inovação Tecnológica: Profa. Dra. Sara Gonçalves Antunes de Souza

Pró-reitor de Pós-graduação: Prof. Dr. André Luiz Sena Guimarães

Coordenadoria de Pós-graduação Lato-sensu: Prof. Dr. Marcos Flávio Silveira Vasconcelos D'Angelo

Coordenadoria de Pós-graduação Stricto-sensu: Prof. Dr. Marcelo Perim Baldo

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS DA SAÚDE

Coordenador(a): Prof. Dr. Alfredo Maurício Batista de Paula

Subcoordenador(a): Prof. Dr. Renato Sobral Monteiro Júnior

UNIVERSIDADE ESTADUAL DE MONTES CLAROS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS DA SAÚDE
MESTRADO E DOUTORADO ACADÊMICO

NOME DO(A) DISCENTE: LAERCIO IVES SANTOS

( ) Mestrado Acadêmico em Ciência Da Saúde

(x) Doutorado Acadêmico em Ciências Da Saúde

TÍTULO DO TRABALHO DE CONCLUSÃO DE CURSO (TCC):
"Adaptação de algoritmos híbridos baseados em Aprendizagem de Máquinas para aplicação em problemas na área de Saúde com base de dados desbalanceadas"

ÁREA DE CONCENTRAÇÃO:	**LINHA DE PESQUISA**:
( ) Mecanismos e aspectos clínicos das doenças	( ) Etiopatogenia e Fisiopatologia das Doenças
	( ) Clínica, Diagnóstico e Terapêutica das Doenças
(X) Saúde coletiva	( ) Educação em Saúde, Avaliação de Programas e Serviços
	(X) Epidemiologia Populacional e Molecular

BANCA (TITULARES)		**ASSINATURAS**
PROF. DR. Marcos Flávio Silveira Vasconcelos D'Angelo	ORIENTADOR	
PROF. DR João Batista Mendes	COORIENTADOR	
PROF. DR. André Luiz Sena Guimarães		
PROF. DR. Marcelo Perim Baldo		
PROF. DR Reinaldo Martinez Palhares		
PROF. DR Victor Hugo Costa de Albuquerque		

BANCA (SUPLENTES)	**ASSINATURAS**
PROFª. DRª. Carla Silvana de Oliveira e Silva	
PROFª. DRª. Desirée Sant'Ana Haikal	

A análise realizada pelos membros examinadores da presente defesa pública de TCC teve como resultado parecer de:

[X] **APROVAÇÃO** [ ] **REPROVAÇÃO**

AGRADECIMENTOS

Gostaria de agradecer primeiramente a Deus pela vida e saúde não só minhas mas de todos que contribuíram para a realização deste trabalho.

Agradeço à minha família por todo o amor, pela educação e pelo apoio incondicional durante toda trajetória acadêmica. Em especial, agradeço a minha esposa Késsia, principalmente, por cuidar de nossa família durante minhas ausências.

A todos os colegas do IFNMG Campus Montes Claros, no qual destaco o professor Lúcio Dutra e a professora Luciana Balieiro, por ouvir, por discutir, por discordar e assim também contribuirem para o amadurecimento de meu pensamento científico.

Aos professores e colegas do PPGCS, aos professores e colegas do PPGMCS, aos professores e colegas do curso Sistemas de Informação da UNIMONTES, aos professores e colegas do IFNMG-Campus Januária e aos professores e colegas da Escola estadual de Cônego Marinho. Aos colegas e professores que atuaram diretamente na construção dos artigos que compõem esta tese. Em fim, muito obrigado a todos que “colocaram pelo menos um tijolo” na construção deste trabalho.

Aos membros da banca de qualificação pelas sugestões realizadas no sentido de melhorar este documento.

E por fim, gostaria de agradecer, especialmente, ao meu orientador, professor Marcos Flávio, e ao meu coorientador, professor João Batista, pela orientação, pelos ensinamentos, por facilitar o enfrentamento de situações complexas e, principalmente, pela amizade recíproca demonstrada durante todos esses anos.

RESUMO

Recentemente, vários pesquisadores têm utilizado modelos desenvolvidos a partir de algoritmos de Aprendizado de Máquina como ferramentas computacionais para auxiliar na resolução de problemas da medicina e da saúde pública. Esses modelos podem auxiliar especialistas na tomada de decisão no que diz respeito a eventos relacionados à área de saúde, melhorando a qualidade do atendimento e possibilitando a redução de gastos. Entretanto, esses pesquisadores também enfrentam alguns desafios e três deles são abordados nesse trabalho: o desbalanceamento de dados; a interpretabilidade dos modelos; e as incertezas de classificação. Dessa forma, esse trabalho apresenta duas abordagens. A primeira abordagem lida com o desbalanceamento de dados e as incertezas de classificação e é aplicada no monitoramento e prevenção de quedas em idosos. Essa abordagem utiliza um algoritmo de Inteligência de Enxames com janelas de pertinência formadas a partir de sinais capitados por dispositivos de Identificação por Rádio Frequência (*RFID*) para monitorar e detectar o movimento de saída de cama dos participantes idosos. A segunda abordagem é utilizada para prever Acidente Vascular Cerebral em um conjunto de dados altamente desbalanceado. Na abordagem, um Sistema Imunológico Artificial trata o desbalanceamento de dados e uma Árvore de Decisão prover um modelo de classificação de fácil compreensão. As duas abordagens têm resultados melhores quando comparadas com abordagens de estado da arte, e isso é um passo importante na direção do desenvolvimento de tecnologias baseadas em Aprendizado de Máquina aplicadas na área de saúde.

Palavras-chave: Aprendizado de Máquina, Saúde Pública, Inteligência Artificial, Prevenção de quedas em idosos, Acidente Vascular Cerebral.

ABSTRACT

Recently, several researchers have used models developed from Machine Learning algorithms as computational tools to help solve problems in medicine and public health. These models can help specialists in decision making regarding events related to healthcare, improving the quality of care and enabling cost reduction. However, these researchers also face some challenges and three of them are addressed in this work: class imbalancing problem; interpretability of models; and classification uncertainties. Thus, this work presents two approachs. The first approach deals with data imbalance and classification uncertainties and is applied in monitoring and preventing falls in the elderly. This approach uses a Swarm Intelligence algorithm with membership windows formed from signals captured by Radio Frequency Identification (RFID) devices to monitor and detect the movement of elderly participants out of bed. The second approach is used to predict Stroke on a highly unbalanced dataset. In the approach, an Artificial Immune System handles data imbalance and a Decision Tree provides an easy-to-understand classification model. Both approaches have better results when compared to state-of-the-art approaches, and this is an important step towards the development of Machine Learning-based technologies applied in healthcare.

Keywords: Machine Learning, Healthcare, Artificial Intelligence, Prevention of falls in the elderly, Stroke.

APRESENTAÇÃO

Em dezembro de 2018, quando ingressei no doutorado em Ciências da Saúde, não imaginava quais transformações essa experiência proporcionaria em minha vida. Minha trajetória como estudante começa na Escola Estadual de Cônego Marinho, passando pelo curso Técnico em Informática realizado no antigo CEFET-Januária, hoje Instituto Federal do Norte de Minas Gerais-Campus Januária e, logo depois, pelo bacharelado em Sistemas de Informação na UNIMONTES.

Meu primeiro contato com a pesquisa científica, não ocorreu na graduação, como com a maioria dos pesquisadores. Em sistemas de informação não tive a oportunidade de realizar iniciação científica ou trabalhar em projetos de pesquisa. Somente, em 2014, no mestrado em Modelagem Computacional e Sistemas, também na UNIMONTES, que tive o primeiro e significativo contato com o método científico. Durante o mestrado realizei os primeiros trabalhos com técnicas de Aprendizado de Máquina e computação Bioinspirada. Em minha dissertação desenvolvi uma abordagem de classificação baseada em Inteligência de Enxames, sob a supervisão dos professores Marcos Flávio (Orientador) e Reinaldo Palhares (coorientador), e apliquei essa abordagem em dois problemas de detecção de falhas.

Acredito que durante o doutorado consegui contribuir em estudos importantes para a área de saúde e também para o campo do Aprendizado de Máquina. No artigo que abordou técnicas de Inteligência Computacional para validação do questionário IPAQ, publicado na *Plos One* e sob a orientação do professor Marcos Flávio, entendi que técnicas baseadas em Aprendizado de Máquina podem ser utilizadas para diversos fins e não só atuar na predição de eventos. Quando os professores André Sena e Marcos Flávio elaboraram o protocolo que originou o artigo intitulado "*Therapeutic perceptions in antisense RNA-mediated gene regulation for COVID-19"* publicado na revista GENE e me convidaram para fazer parte da equipe de pesquisa, eu, sinceramente, não acreditava que seria possível elaborar em tempo hábil a abordagem de busca por alvos terapêuticos que ambos propunham, afinal pacientes acometidos pela *COVID-19* não podiam esperar. Entretanto, a oportunidade de contribuir de alguma maneira no combate dessa doença era desafiador e intrigante. Destaco aqui também o trabalho "*Development of a two-year death prediction model among patients with Chagas Disease using methods based on machine learning",* submetido à revista *Plos NTD*. Agradeço especialmente à professora Desirrê Sant'ana e à Ariela Mota por permitirem minha contribuição nesse estudo e espero que ele possa ser importante no sentido de melhorar as condições de saúde da população acometida pela doença de Chagas.

Ao participar de equipes de pesquisa tão heterogênicas, pude compreender a importância da interdisciplinaridade na resolução de problemas complexos e também deste programa de pós-graduação no cenário científico e para a sociedade. Portanto, realizar trabalhos, que têm como objetivo proporcionar respostas consistentes para problemas complexos, é menos dispendioso quando os esforços são somados por uma equipe competente e com conhecimento diversificado. Como principal desafio dessa fase, eu destaco a dificuldade em desenvolver essa tese como algo condizente com os anseios deste programa e das Ciências da Saúde em geral.

Esta tese segue a formatação preconizada pelo PPGCS - Unimontes, que recomenda a apresentação de uma primeira seção com a introdução/revisão da literatura e os objetivos do trabalho. Uma segunda seção, que consiste na apresentação dos dois artigos desenvolvidos no âmbito do doutorado. E, por fim, uma seção de considerações finais.

Em relação aos dois artigos desta tese, o primeiro utiliza uma abordagem baseada em Inteligência de Enxames e Conjuntos Difusos para detectar saída da cama de pacientes idosos e no segundo propomos uma abordagem de Árvores de Decisão induzidas por Programação Genética (PG) para classificar dados de Acidente Vascular Cerebral (AVC). Propomos, ainda no segundo artigo, um mecanismo para balancear as classes do conjunto de dados de AVC já que a classe positiva era representada por somente 1.89% dos dados do conjunto.

No primeiro artigo, adaptamos a abordagem proposta na dissertação de mestrado inserindo o conceito de "Janelas de Pertinência *Fuzzy*" na tomada de decisão na emissão do alerta da saída da cama. A proposta da janela seria utilizar o conceito de classificação por agregação, como a realizada em *ensembles* como *Random Forest*, por exemplo. Pois é sabido que esse tipo de classificação, por vezes, superam a assertividade dos modelos onde apenas um classificador é utilizado. A principal diferente de nossa abordagem para as abordagens do tipo *ensemble,* está no fato de nossa abordagem utilizar janelas deslizantes no tempo e um único classificador, ao passo que os *ensembles* utilizam diversos classificadores em um único ponto no tempo.

Em relação ao segundo artigo é importante dizer que a proposta de Árvores de Decisão induzidas por PG surgiu durante a realização do estudo de predição de mortes por doença de Chagas. Pois este estudo utilizava um conjunto de dados com mais de 80 características, inicialmente. A intenção era utilizar a PG para reduzir a quantidade de características durante a indução da população de árvores. Posteriormente, o professor Marcos Flávio e o professor João Batista propuseram utilizar a PG no problema de predição de AVC e

para tanto deveríamos propor um mecanismo que pudesse realizar o balanceamento das classes.

Acredito que esta tese poderá inspirar docentes e discentes do PPGCS e de outros programas de pós-graduação a realizarem pesquisas que tem como objetivo o desenvolvimento de técnicas e ferramentas de Aprendizado de Máquina e computacionais e aplicá-las na resolução de problemas da área de saúde, e assim, ajudar na melhoraria das condições de saúde e da qualidade de vida das populações estudadas.

SUMÁRIO

1-Introdução

1.1 – Considerações iniciais

O direito à saúde no Brasil, garantido pela Constituição Federal de 1988, deve ser garantido de forma integral e universalizada. Apesar dessa garantia, a população brasileira enfrenta muitos desafios para ter a saúde assegurada pelo Estado na amplitude do seu conteúdo, pois a garantia desse direito só é possível a partir de ações em áreas diversas e complementares (BRITO-SILVA et al., 2012). Ao longo dos anos, avanços importantes devem ser considerados. Por exemplo, pode-se citar a redução da mortalidade por doenças transmissíveis e por causas evitáveis da morbimortalidade materno-infantil entre os anos 1990 e 2015, o aumento da expectativa de vida da população que passou de 68,4 anos em 1990 para 75,2 anos em 2016 (SOUZA et al., 2018) e a erradicação de doenças como a Poliomielite que teve como principal fator de contribuição a vacinação (VERANI & LAENDER, 2020). De outro modo, alguns desafios ainda persistem, por exemplo o acesso à saúde, que impacta direta ou indiretamente a mortalidade e a expectativa de vida, e ainda é um obstáculo para populações rurais em comparação com quem vive em áreas urbanas (ARRUDA, et al.,2018; CROSS et al., 2020; KIRBY & YABROFF, 2020). Além disso, é importante mencionar a falta de profissionais especializados e os altos custos em determinadas áreas da saúde.

Na tentativa de lidar com esses desafios, pesquisadores têm aplicado algoritmos de Aprendizado de Máquina (AM) na área de saúde. O Aprendizado de Máquina pode ser definido como um conjunto de ferramentas e métodos para identificar padrões em dados. Esses padrões podem ser usados para aumentar nossa compreensão de um problema específico (por exemplo, identificar fatores de risco de uma determinada doença) ou fazer previsões sobre o futuro (por exemplo, prever se um paciente vai contrair uma determinada doença) (WIENS & SHENOY, 2018).

Algoritmos de AM analisam grandes bases de dados com um tempo relativamente baixo, podem tratar relações complexas entre os dados, o que os tornam tão ou mais precisos que especialistas humanos em algumas situações (DEO, 2015). A utilização de técnicas de AM tem o potencial demonstrado para melhorar os indicadores gerais de saúde e de qualidade de vida, reduzir os gastos com a saúde de algumas populações e avançar na pesquisa clínica (WARING et al., 2020). Outro aspecto importante que tem contribuído para o aumento da utilização de AM na área de saúde, foi a disponibilidade de grandes quantidades de dados de alta qualidade sobre pacientes e instalações, tendo em vista que esses algoritmos necessitam

de um volume de dados considerável para funcionarem de forma eficiente (WIENS & SHENOY, 2018).

AM baseia-se em conceitos de vários campos, incluindo ciência da computação, estatística e otimização. Em geral, a maioria dos problemas, que se pode aplicar o AM, podem ser formulados como um problema de otimização em relação a um conjunto de dados. Nessas configurações, o objetivo é encontrar um modelo que melhor explique o conjunto de dados, pode-se dizer que o algoritmo aprende a partir dos dados (WIENS & SHENOY, 2018). Essa tarefa é denominada de aprendizado e ela pode ocorrer basicamente de 3 formas: Aprendizado Supervisionado, Aprendizado Não Supervisionado e Aprendizado por Reforço.

No Aprendizado Supervisionado, instâncias de dados são rotuladas previamente (conjunto de treinamento) de acordo com uma característica específica de interesse (pacientes doentes ou pacientes saudáveis, por exemplo). O processo de treinamento busca por uma função que mapeia um conjunto de variáveis preditoras para a variável de interesse previamente rotulada (também denominada de variável desfecho ou classe) (WIENS & SHENOY, 2018). Diferente do Aprendizado Supervisionado, no Aprendizado Não Supervisionado as instâncias de dados não estão previamente rotuladas. Nesse tipo de aprendizado, o objetivo é mensurar as similaridades entre as instâncias analisadas e assim agrupar tais instâncias (CELEBI & AYDIN, 2016). Por fim, no Aprendizado por Reforço, modelos de aprendizado de máquina tentam tomar decisões em sequência para atingir uma determinada meta através de vários movimentos de tentativa e erro. Movimentos considerados corretos recebem reforços positivos e movimentos considerados incorretos recebem reforço negativo (KAELBLING et al., 1996).

As principais contribuições dessa Tese estão relatadas ao longo de 2 artigos. O primeiro artigo trata o problema de monitoramento de saída do leito de pessoas idosas, e, nesse contexto, propomos a utilização de janelas de pertinência na tomada de decisão como contribuição. A utilização das janelas demostrou ser eficiente em amenizar as incertezas de classificação, inerentes ao problema e causadas, principalmente, pela perda de sinais. No segundo artigo, propomos uma abordagem baseada em um Sistema Imunológico Artificial para lidar com o desbalanceamento de dados e, ainda, inserimos e avaliamos um operador de simplificação no mecanismo de indução de árvores de decisão via Programação Genética. A abordagem se mostrou eficiente tanto na melhoria da assertividade do modelo final, quanto na indução de árvores menores e mais interpretáveis.

1.2 – Algoritmos de Classificação na área de Saúde

A Classificação é uma tarefa do tipo Aprendizado Supervisionado, e, portanto, seu objetivo é atribuir categorias ou classes, predefinidas, a instâncias de dados. Por exemplo, um sistema baseado em técnicas de AM pode ser utilizado para classificar pacientes como diabéticos ou não diabéticos (SISODIA et al, 2018). Essa decisão pode ser tomada utilizando informações do paciente como: idade, índice de massa corporal, pressão arterial diastólica, glicose, dentre outras. O sistema aprenderá a diferenciar pacientes saudáveis de pacientes diabéticos.

Separar objetos em classes é uma tarefa que o ser humano, frequentemente, executa facilmente. Recebemos sinais externos através de nossos sentidos e podemos reconhecer, em algum contexto, a que classe ou categoria pertencem estes sinais. Podemos fazer isto quase que imediatamente e com pouco esforço, caso o conhecimento necessário já tenha sido adquirido por meio de algum processo de aprendizagem (SEMOLINI, 2002). De maneira parecida, a predição realizada por um algoritmo de classificação acontece em duas etapas. Na primeira etapa um modelo é construído para descrever um conjunto pré-determinado de classes. Tal construção é feita analisando a base de dados, onde as instâncias são descritas por atributos e cada uma delas pertence a uma classe predefinida, identificada por um dos atributos, chamado atributo classe. Esse conjunto de instâncias é chamado conjunto de treinamento. Assim, a primeira etapa também é chamada de Treinamento ou Aprendizado. Na segunda etapa (Predição), o modelo construído é utilizado para classificar um conjunto de instâncias diferente do conjunto utilizado na primeira etapa. O propósito de um modelo de classificação é predizer a classe de instâncias quando essa é desconhecida. Ou seja, dada uma instância para um problema qualquer predizer a qual classe essa instância pertence. Esse processo é ilustrado pela Figura 1 em duas etapas (treinamento e predição). Na etapa de treinamento, um conjunto de dados rotulado é apresentado ao algoritmo de AM. O algoritmo busca por um modelo preditivo que será utilizado para identificar as classes de instâncias de um conjunto de dados não rotulado.

No cotexto da área de Saúde, modelos desenvolvidos a partir de algoritmos de classificação podem ser utilizados de três maneiras distintas: Predizer o risco de um evento ocorrer dado que ele já ocorreu (diagnóstico);- ou ainda não ocorreu (prognóstico);- e ainda, a predição pode ser direcionada como uma ação preventiva. No primeiro caso, ferramentas computacionais desenvolvidas a partir de modelos de AM podem ajudar médicos e profissionais de saúde a diagnosticar condições de saúde complexas. Os algoritmos são alimentados com sinais de entrada (por exemplo, dados extraídos de imagens de ressonância

magnética) e, a partir de tais entradas, determinam uma condição de saúde. Além disso, a escassez de médicos especialistas, principalmente em áreas rurais, aumenta o problema do diagnóstico precoce e preciso de algumas doenças (como câncer de mama, por exemplo), causando maior taxa de mortalidade. Nesses casos, essas ferramentas podem ser utilizadas em análises prévias e ajudar na tomada de decisão, indicando se um determinado paciente precisa ou não ser avaliado por um especialista (YASSIN et al., 2018).

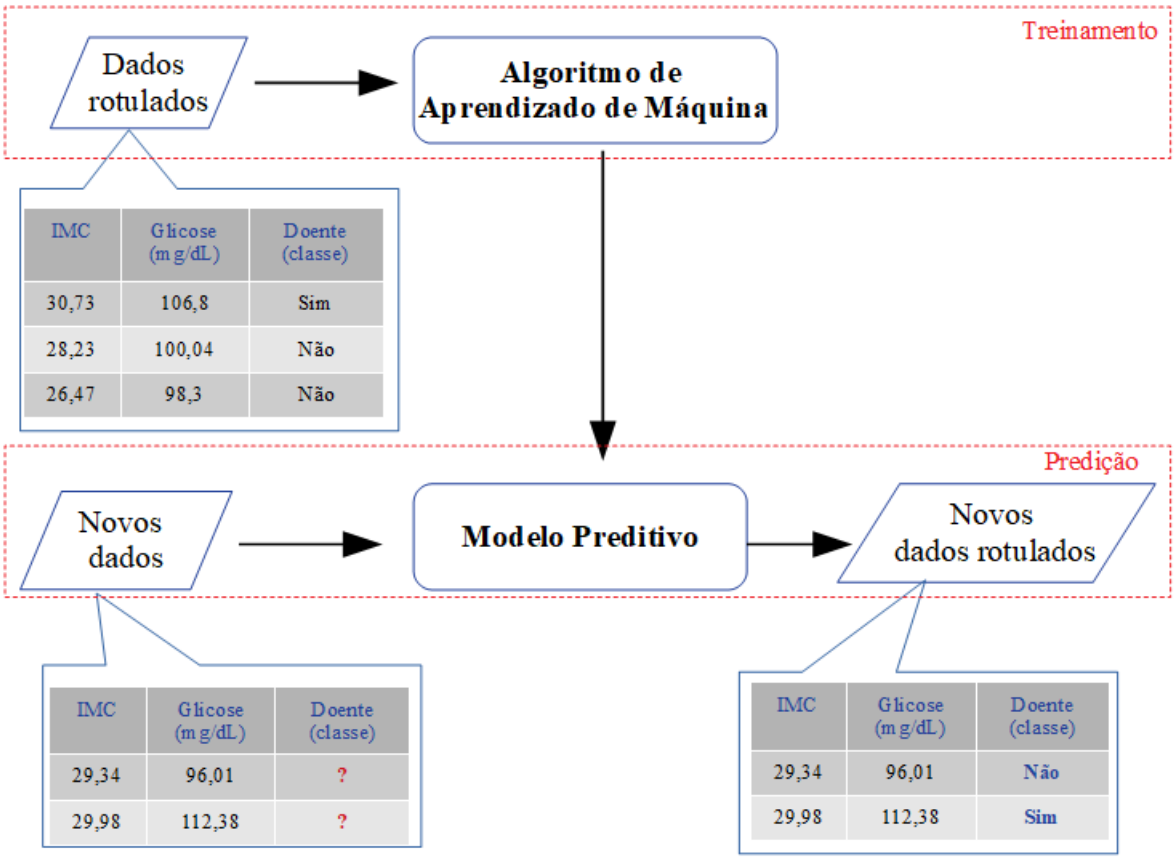


Figura 1: Esquema para aplicação de algoritmos de Aprendizado de Máquina em problemas de Classificação

No segundo caso, os algoritmos de classificação podem ser utilizados para predizer o risco de indivíduos ou populações em relação a um determinado desfecho e essa informação pode ser útil para a comunicação entre profissionais de saúde e posterior estabelecimento de condutas visando uma melhora prognóstica. No terceiro caso, um conjunto de ações preventivas podem ocorrer em diferentes momentos da história do desfecho de interesse, que pode ser: antes da instalação dos fatores de risco, antes do diagnóstico clínico ou antes das complicações desse desfecho. Nesse caso, os algoritmos de AM podem auxiliar equipes de saúde a identificar indivíduos com alto risco de desenvolver um desfecho específico,

oportunizando a essas equipes o planejamento e execução de intervenções que visam reduzir o risco do desfecho nesses indivíduos (SANTOS, 2018).

Como exemplo do primeiro caso, pode-se citar o estudo realizado por ZOABI et al. (2021) que utiliza modelos de AM para auxiliar no diagnóstico da COVID-19. O objetivo do estudo é criar um sistema capaz de auxiliar equipes médicas a realizar o diagnóstico da COVID-19, principalmente em regiões onde há limitação de recursos, como em áreas rurais. No estudo, os algoritmos de AM foram treinados utilizando variáveis derivadas de uma anamnese referente à doença e 51.831 indivíduos testados.

De outra forma, em (SANTOS et al., 2019) foram utilizados algoritmos de AM para predizer o risco de morte em 5 anos de 2.808 idosos a partir de 37 variáveis preditoras relacionadas ao perfil demográfico, socioeconômico e de saúde dos participantes. Esses modelos podem, por exemplo, ajudar a identificar pacientes com baixo risco de morte e direcioná-los a serviços de saúde de menor complexidade e, assim, reduzir os gastos hospitalares com essa população. De outra forma, os pacientes com risco de morte alto poderiam ser direcionados a serviços de saúde mais específicos e com intervenções personalizadas e isso poderia reduzir esse risco.

Por fim, como exemplo de estudos que utilizaram algoritmos de classificação para estabelecer ações preventivas, é apresentado o estudo realizado por PAN et al. (2017), em que, dados de 6457 nascimentos em Illinois (EUA) e 17 variáveis preditoras foram utilizados para prever desfechos desfavoráveis em crianças no momento do nascimento ou morte no primeiro ano de vida. O objetivo principal da análise preditiva foi estabelecer critérios para selecionar as gestantes para um programa de acompanhamento especializado.

A tabela 1 apresenta outros estudos sobre aplicação de algoritmos baseados em AM em problemas da área de Saúde.

Mesmo com a ampla utilização de algoritmos baseados em AM na área de saúde, alguns desafios ainda persistem. Portanto, a próxima seção aborda os principais desafios e discutir brevemente algumas oportunidades científicas e financeiras relacionadas ao assunto.

1.3 – Desafios, Limitações e Oportunidades

A performance de modelos de AM, em geral, é diretamente influenciada pela qualidade dos conjuntos de dados utilizados na etapa de treinamento. Uma característica inerente nos conjuntos de dados gerados a partir de problemas na área de Saúde é o desbalanceamento de dados. O desbalanceamento de dados está relacionado ao fato de que alguns eventos raros, catalogados em conjuntos de dados, resultam em classes que não são igualmente representadas. Por exemplo, em um conjunto de dados sobre possíveis pacientes

com câncer, a quantidade de pacientes saudáveis (instâncias negativas) é significativamente maior que a quantidade de pacientes diagnosticados com câncer (instâncias positivas). A classe com maior prevalência é chamada classe majoritária, enquanto a classe mais rara é chamada classe minoritária (LI et al., 2016).

Tabela1: Aplicação de algoritmos de AM em problemas da área de saúde.

Propósito	**Estudo**
Diagnosticar Câncer de pele	(VIDYA e KARKI, 2020), (ROFFMAN et al, 2018), (ESTEVA, 2017)
Diagnosticar Câncer de mama	(PAL, 2018), (AMRANE, Meriem et al, 2018), (ABDEL-ZAHER e ELDEIB, 2016), (BHARDWAJ e TIWARI, 2015)
Diagnosticar Câncer de pumão	(YU et al, 2019), (RATTAN, 2017)
Diagnosticar câncer de Diabetes	(KAUR e CHHABRA, 2014), (CHRISTINA e SANTIAGO, 2018)
Diagnosticar doenças cardíacas	(JOSHI, DANGRA e RAWAT, 2018) e Dengue (GAMBHIR, 2018)
Identificar indivíduos com risco suicida	(DESMET e HOSTE, 2018), (OH, 2017), (JI, 2018)
Segurança alimentar nutricional	(SILVA, 2015)
Identificar indivíduos com transtorno Bipolar	(PONTE, 2018)
Buscar por doadores de sangue	(SILVA, 2018)

Os algoritmos de AM têm dificuldade em tratar dados desbalanceados por tenderem a classificar todas as instâncias na classe majoritária em detrimento a classe minoritária que, em geral, se caracteriza como o evento de interesse (LI et al., 2017). Para ilustrar o problema, vamos supor a utilização de um algoritmo de AM para diagnosticar uma doença em um conjunto de dados no qual pacientes com diagnóstico positivo são somente 1% do total de pacientes analisados. Nesse cenário hipotético, os algoritmos de AM tendem a classificar todos os pacientes como “não doentes”, ou seja, com o diagnóstico negativo e, dessa forma, obter uma assertividade de 99%. Note que mesmo com uma alta assertividade, o algoritmo de AM errou em todos os pacientes com diagnóstico positivo. Ou seja, o algoritmo não foi capaz de detectar o evento de interesse.

A Figura 2 ilustra 4 situações relacionadas ao desbalanceamento de dados. Primeiramente, a Figura 2 A) exemplifica uma situação na qual o desbalanceamento de dados ocorre, porém, as classes são linearmente separáveis. Quando isso ocorre, o desbalanceamento de dados por si só não dificulta as ações do algoritmo de AM. Entretanto, quando aliado ao desbalanceamento de dados, há uma sobreposição entre as classes (situação ilustrada pela

parte Figura 2 B), as chances de um exemplo da classe minoritária ser corretamente classificado diminui.

Outra característica relevante é o tamanho da amostra, pois amostras pequenas (Figura 2 C) tendem a não representar o espaço de aplicação do algoritmo de AM. Particularmente, quando há desbalanceamento de dados, a classe minoritária é a mais prejudicada por esse tipo de amostragem. Quando mais instâncias de dados podem ser usadas, relativamente, mais informações sobre a classe minoritária beneficiam a classificação (SUN et al., 2007). Por fim, a Figura 2 D) ilustra o desbalanceamento de dados com mais de duas classes. Estudos anteriores mostram que a dificuldade de desenvolver modelos acurados é diretamente proporcional à quantidade de classes, pois, as relações entre as classes não são mais óbvias. Uma classe pode ser majoritária quando comparada a algumas outras classes, mas minoritária ou bem equilibrada para as demais (SÁEZ et al., 2016).

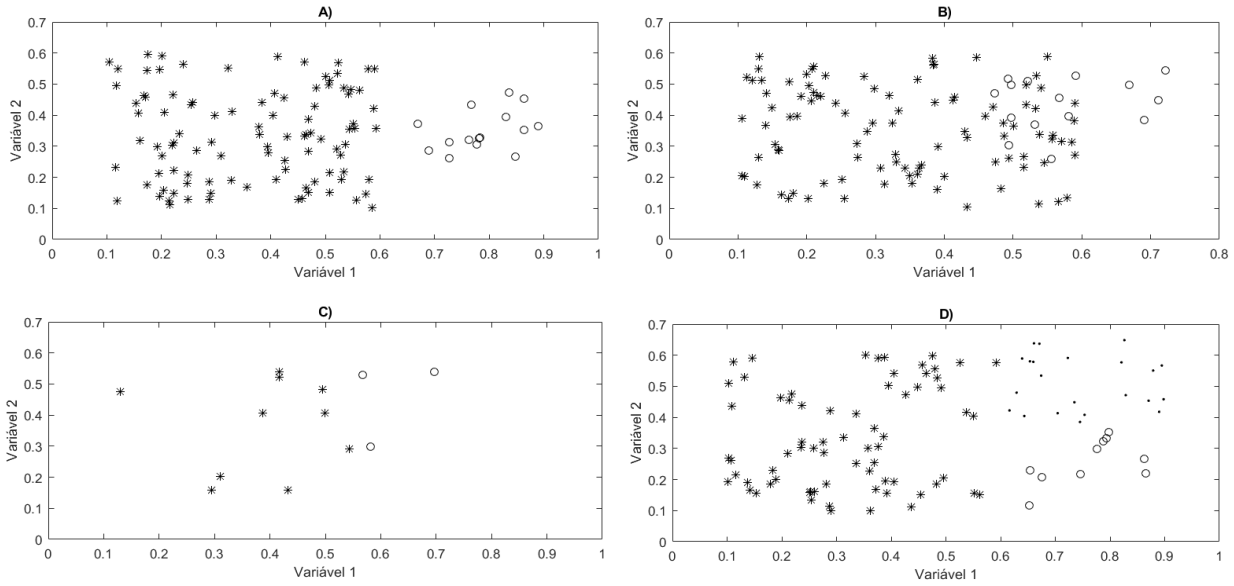


Figura 2: Exemplos de classes desbalanceadas. A) Duas classes linearmente separáveis; B) Com sobreposição entre as classes; C) Amostra pequena; e, D) Mais de duas classes (multi classe)

O desbalanceamento de dados ainda é um problema em aberto. Primeiramente por que as principais estratégias como *Over-sampling* (replicar exemplos da classe minoritária), *Under-sampling* (eliminar exemplos da classe majoritária) ou criação de algoritmos de aprendizado sensíveis ao custo, sofrem de problemas como sobreajuste aos dados de treinamento e perda de informações relevantes (KAUR et al., 2019). Além disso, em conjuntos de dados de domínios como medicina e saúde pública, o desbalanceamento de dados deve ser tratado cuidadosamente, pois a classe minoritária, em geral, se refere ao desfecho adverso e desfavorável (como morte, doença, alto risco de internação e etc). Por

exemplo, classificar um paciente enfermo como saudável inviabiliza a intervenção e tratamento do mesmo e dependendo da gravidade da patologia, pode levar o paciente a morte.

A ausência de dados (*missing data*, em inglês) é outro desafio inerente da aplicação de algoritmos de AM em diversos domínios. Na área da Saúde isso ainda é um problema que desperta o interesse de pesquisadores por dois motivos. Primeiramente, as duas principais abordagens para tratar dados ausentes (apagar todas as instâncias que tenham ao menos uma variável com valor ausente; substituir os valores ausentes artificialmente) sofrem de problemas como perda de informação ou criação de informações mentirosas e enviesadas (CISMONDI et al., 2013). Por outro lado, quando a ausência de dados ocorre de forma não aleatória, os algoritmos podem interpretar incorretamente os dados disponíveis e, consequentemente, não oferecer benefícios para pacientes cujos dados estão faltando no conjunto de treinamento. Por exemplo, pacientes com baixo nível socioeconômico, no geral, têm acesso a uma menor quantidade de testes diagnósticos e medicamentos para doenças crônicas e acesso limitado aos sistemas de saúde. Assim, esses pacientes podem ter informações insuficientes nos registros eletrônicos. A consequência disso, é que os algoritmos de AM não serão treinados adequadamente nesse subgrupo o que comprometeria seus resultados. Além disso, pacientes desse subgrupo poderiam ser privados de ações preventivas e personalizadas desenvolvidas a partir das previsões dos modelos treinados (GIANFRANCESCO et al., 2018).

Em relação a variedade de algoritmos de classificação é importante destacar que métodos robustos do tipo caixa-preta como Redes Neurais Artificiais (DA SILVA, 2017) e Máquinas de Vetores de Suporte (HEARST, 1998) não são intuitivamente interpretáveis (CARUANA et al., 2015). A interpretabilidade, está associada a capacidade de justificar ou apresentar em termos compreensíveis uma predição realizada pelo modelo no conjunto de dados, possibilitando explorar um dado problema, gerar novas ideias sobre como resolvê-lo e melhorar a compreensão dos especialistas (DOSHI-VELEZ & KIM, 2017). No contexto da prática médica, a interpretabilidade de um modelo desenvolvido a partir de um algoritmo de AM tem sido geralmente considerada crítica para estabelecer a confiança dos médicos nessas ferramentas e consequentemente possibilitar a validação de seus resultados. Assim, prestadores de serviços clínicos e outros tomadores de decisão na área da Saúde observam a interpretabilidade das previsões do modelo desenvolvido a partir de técnicas de AM como uma prioridade para implementação e utilização (AHMAD et al., 2018).

Modelos de predição desenvolvidos a partir de classificadores do tipo Árvore de Decisão são intuitivamente interpretáveis quando sua complexidade é baixa (há poucos nós de

decisão). Uma Árvore de Decisão é estrutura definida recursivamente e composta por nós de decisão e nós folhas. Um nó de decisão contém um teste sobre alguma variável preditora e para cada resultado desse teste existe uma aresta para uma subárvore ou nó folha. Já um nó folha corresponde a uma das classes do problema. A Figura 3 A) ilustra um modelo derivado de Árvore de Decisão induzido a partir de instâncias de dados fictícias e a Figura 3 B) ilustra tais instâncias divididas em 4 quadrantes (1º, 2º, 3º e 4º). Com o objetivo de exemplificar como a predição realizada por esse tipo de modelo pode ser facilmente compreendida, vamos utilizar a regra "Idade >= 72 **E** IMC >=28 ENTÃO Classe=1", que pode ser derivada do modelo da Figura 3 A). As instâncias cobertas por essa regra estão representadas no 1º quadrante do gráfico na Figura 3 B) e a partir dessa regra é possível concluir que um indivíduo com 72 anos ou mais e com um índice de massa corporal (IMC) a partir de 28, tem alta probabilidade de estar doente.

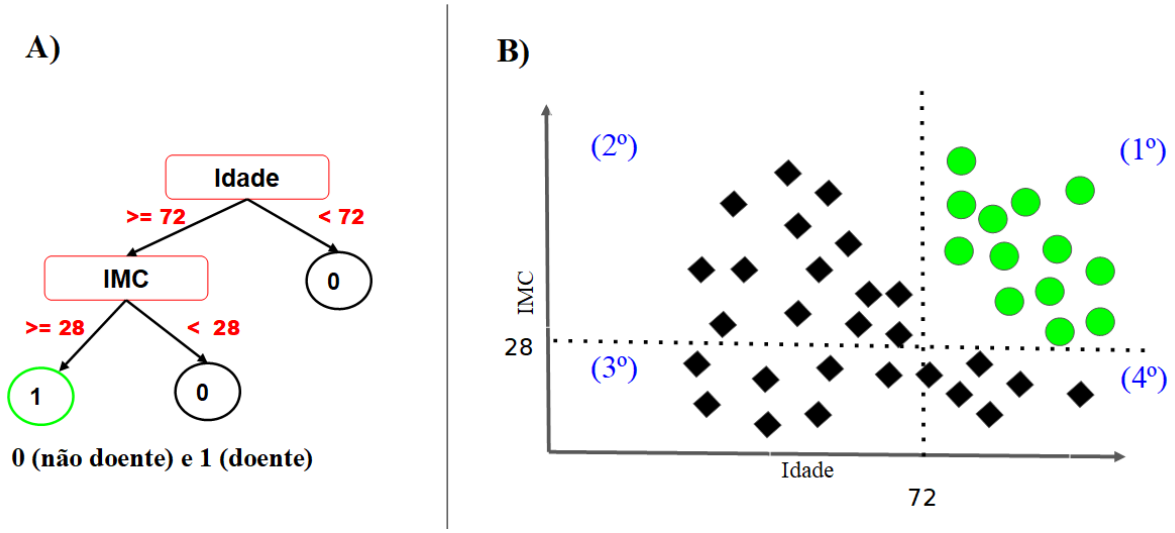


Figura 3: Classificação utilizando um modelo baseado em Árvore de Decisão fictício.

Na literatura há relatos de abordagens que utilizam explicadores para lidar com o problema de interpretabilidade de modelos do tipo caixa-preta. A utilização desse tipo de abordagem tem crescido entre desenvolvedores de modelos de AM. Porém, a explicabilidade que elas fornecem não é totalmente fiel ao modelo de predição original. Resumidamente, elas aproximam o modelo caixa-preta de um modelo linear intuitivamente interpretável (como um sistema baseado em regras ou Árvore de Decisão, por exemplo). Apesar disso, esses explicadores são considerados abordagens interessantes quando aplicados em problemas de classificação em conjuntos de dados complexos (como textos e imagens médicas). Nesse tipo de problema, os explicadores externos conseguem indicar aos especialistas humanos os

principais elementos utilizados pelo modelo para realizar a predição e assim fornecer algum grau de interpretabilidade (SELVARAJU et al., 2017; RIBEIRO et all., 2016).

Um quarto desafio está relacionado às incertezas inerentes na tomada de decisão em saúde pública e medicina em geral. Esse conhecimento impreciso é inevitável e decorrente de diversos fatores, como: Compreensão incompleta dos mecanismos biológicos; Medições imprecisas; Presença simultânea de mais de uma condição e Informações ausentes em uma grande porcentagem de casos (YARDIMCI, 2009). Além disso, o diagnóstico médico envolve tanto fatores objetivos quanto subjetivos e cada paciente pode apresentar diferentes graus de suspeita para várias condições de saúde. Dessa forma, o diagnóstico sempre é feito com algum grau de incerteza (AHMADI et al., 2018). No contexto da predição de eventos relacionados à saúde esses fatos são relevantes, pois, a maioria dos problemas de classificação são formulados de forma binária, ou seja, a instância de dados pertence ou não pertence a uma determinada classe. Por exemplo, em um problema de classificação de uma doença como câncer de mama, o algoritmo de classificação vai determinar se um paciente está doente ou não está doente. Entretanto, as incertezas inerentes ao problema podem prejudicar o processo de classificação. Nesses casos, técnicas que tratem essas incertezas podem ser mais adequadas que métodos clássicos.

Incertezas na classificação ocorrem no conjunto de dados utilizado no trabalho desenvolvido em (TORRES et al., 2013), que teve como objetivo propor uma metodologia para detectar a saída de idosos da cama em hospitais ou residências. Essa detecção possibilita a intervenção por parte de cuidadores ou enfermeiros e pode reduzir as quedas dessa população nesses ambientes. Na metodologia proposta, dispositivos de identificação por rádio frequência (*RFID*) são presos às roupas dos idosos e são energizados por antenas instaladas nos quartos. Os sinais emitidos são então processados por uma abordagem de Aprendizado de Máquina para tentar identificar o movimento de saída de cama. As incertezas na classificação ocorrem principalmente devido a dois motivos, ilustrados na Figura 4. Primeiramente, há uma dificuldade em diferenciar o momento em que o idoso está sentado na cama ou na cadeira (movimento 2 na Figura 4 A) do momento que ele se coloca de pé (movimento 4 na Figura 4 A). Isso por que em ambos os movimentos o tronco se posiciona de forma ereta e vertical. O segundo motivo está ilustrado na Figura 4 B), no qual, a antena não consegue energizar corretamente o dispositivo *RFID*, uma vez que o corpo do paciente se coloca entre o dispositivo e a antena.

Por fim, apesar de existirem diferentes abordagens para realizar a classificação, um único método pode não ser adequado para atender todos os requisitos de uma aplicação

(VENKATASUBRAMANIAN, 2005) devido à qualidade dos resultados desses métodos depender das informações fornecidas. O sucesso de um modelo de classificação depende do conjunto de instâncias de treinamento. Assim, um modelo que é adequado para um domínio pode não ser adequado para outro. Uma forma de melhorar o sistema de classificação é utilizar modelos que integrem características de mais de um método, os chamados modelos Híbridos, superando assim as limitações existentes nas estratégias quando usadas de forma individual.

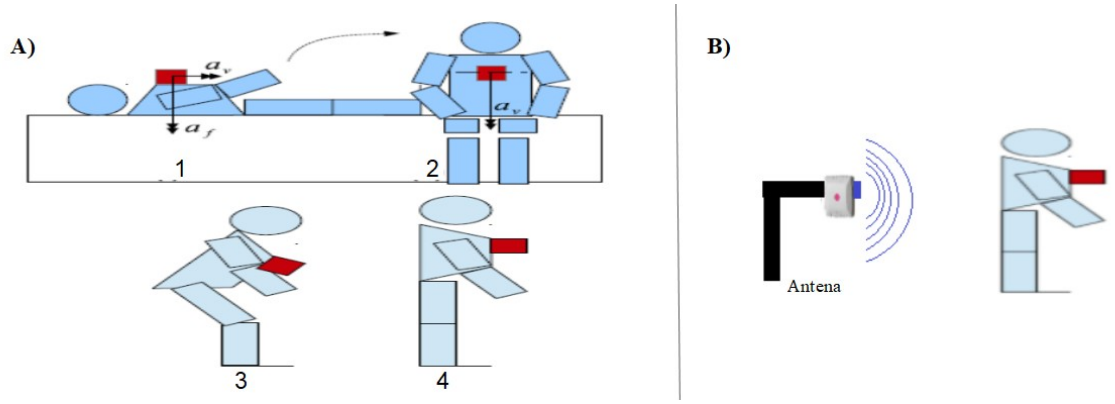


Figura 4: Exemplos de causas de incertezas no problema de monitoramento de saída de cama. A) Adaptada de (TORRES et al., 2013), representa 4 movimentos diferentes que podem ser realizados pelo paciente. B) Ilustra o movimento do paciente quando seu corpo se coloca entre a antena e o sensor (em vermelho).

Em relação às oportunidades inerentes da aplicação de algoritmos de Aprendizado de Máquina na área de saúde, destacamos que, na prática clínica, um sistema baseado em AM poderia atuar na segmentação de pacientes, diminuindo a carga sobre o sistema de saúde e direcionando recursos para os pacientes com maior probabilidade de ter uma necessidade médica real. Esses sistemas também podem funcionar em um cenário de substituição do profissional de saúde. Embora seja improvável que substitua inteiramente os profissionais de saúde, o AM pode realizar certas tarefas com maior consistência, velocidade e reprodutibilidade do que os humanos. Ao automatizar tarefas que não são teoricamente complexas, mas podem ser trabalhosas e demoradas, os profissionais de saúde podem ser liberados para lidar com tarefas mais complexas, que exigem um uso aprimorado do capital humano (HE et al., 2019).

A utilização de tecnologias baseadas em AM tem potencial para melhorar a qualidade dos serviços prestados pelos profissionais de saúde. Essa melhoria é benéfica, para o profissional de saúde em si, para estabelecimento que presta aquele serviço e, principalmente, para os pacientes e usuários dos serviços. Do ponto de vista econômico, a implementação dessas tecnologias na área da saúde poderá, em um futuro próximo, aumentar a demanda por

profissionais cujas habilidades ainda não existem, gerando assim novos empregos. Pesquisas em andamento serão necessárias para desenvolver novas abordagens para aplicações médicas e superar as limitações das abordagens atuais. Portanto, esses benefícios aliados às oportunidades científicas e econômicas justificam a realização deste trabalho.

1.4 Objetivos

1.4.1 – Objetivo Geral

Propor e avaliar adaptações de algoritmos híbridos baseados em aprendizagem de máquinas para classificação de dados em problemas de saúde.

1.4.2 – Objetivos Específicos

Desenvolver uma abordagem baseada em Inteligência de enxames e Conjuntos difusos para monitoramento e emissão de alertas de saída de idosos do leito.

Apresentar uma abordagem que utiliza Sistemas Imunológicos Artificiais e Árvores de Decisão induzidas por Programação Genética para predição de Acidente Vascular Cerebral.

Avaliar as abordagens comparando seus resultados com os resultados de abordagens propostas na Literatura.

1.5 – Artigos desenvolvidos nessa Tese

1. SANTOS, Laércio Ives; D'ANGELO, Marcos Flávio Silveira Vasconcelos; COSME, Luciana Balieiro; DE OLIVEIRA, Heveraldo Rodrigues; MENDES, João Batista; EKEL, Petr Ya. Swarm intelligence and fuzzy sets for bed exit detection of elderly. ***Journal of Intelligent & Fuzzy Systems***, v. 39, n. 1, p. 1061-1072, 2020. DOI: 10.3233/JIFS-191971.

2. SANTOS, Laércio Ives; CAMARGOS, Murilo Osorio; D'ANGELO, Marcos Flávio Silveira Vasconcelos; MENDES, João Batista; DE MEDEIROS, Egydio Emiliano Camargos, GUIMARÃES, André Luiz Sena; PALHARES, Reinaldo Martínez . *Decision Tree and Artificial Immune Systems for Stroke Prediction in Imbalanced Data.* ***Expert Systems With Applications***. v. 191, p. 116221, 2022. DOI: 10.1016/j.eswa.2021.116221.

1.6 - Outros artigos publicados e submetidos no decorrer dessa Tese

1.6.1-Artigos publicados e submetidos em Periódicos

1. SANCHES, Gabriela Luize Guimarães; MENEZES, Agna Soares da Silva; SANTOS, Laércio Ives, DURÃES, Cristina Paixão, FONSECA, Larissa Lopes, BALDO, Marcelo Perim, FARIA, Thais de Oliveira, ANDRADE, Luciano Alves de Araújo, EKEL, Petr Iakovlevitch, SANTOS, Sérgio Henrique Sousa, DE PAULA, Alfredo Maurício Batista, FARIAS, Lucyana Conceição, D'ANGELO, Marcos Flávio Silveira Vasconcelos & GUIMARÃES, André Luiz Sena.. *Local tissue electrical parameters predict oral mucositis in HNSCC patients: A diagnostic accuracy double-blind, randomized controlled trial.* ***Scientific reports***, v. 10, n. 1, p. 1-10, 2020. DOI:10.1038/s41598-020-66351-9
2. FERREIRA FREITAS, Ronilson; SANTOS BRANT ROCHA, Josiane; SANTOS, Laércio Ives; …, D'ANGELO, Marcos Flávio Silveira Vasconcelos . *Validity and precision of the International Physical Activity Questionnaire for climacteric women using computational intelligence techniques***. *PloS one,*** v. 16, n. 1, p. e0245240, 2021. DOI: 10.1371/journal.pone.0245240
3. DE JESUS, Sabrina Ferreira; SANTOS, Laércio Ives; NETO, João Felicio Rodrigues; VIEIRA, Thallyta Maria; MENDES, João Batista; D'ANGELO, Marcos Flavio Silveira Vasconcelos & GUIMARAES, André Luiz Sena. *Therapeutic perceptions in antisense RNA-mediated gene regulation for COVID-19*. **Gene**, v. 800, p. 145839, 2021. DOI: 10.1016/j.gene.2021.145839.
4. DOS SANTOS, Otil Carlos Dias ; SANTOS, Laércio Ives; MACEDO, Reginaldo Morais; D'ANGELO, Marcos Flávio Silveira Vasconcelos. Uma abordagem baseada em Inteligência Computacional para análise dos gastos da saúde dos municípios brasileiros. Submetido à revista **Saúde e Sociedade**.
5. FERREIRA, Ariela Mota; SANTOS, Laércio Ives; SABINO, Ester Cerdeira; RIBEIRO, Antônio Luiz Pinho; SILVA, Léa de Oliveira Campos; DAMASCENO Renata Fiúza; D'ANGELOS, Marcos Flávio Silveira Vasconcelos; NUNES, Maria Do Carmo Pereira; HAIKAL, Desirée Sant´Ana. *Development of a two-year death prediction model among patients with Chagas Disease using methods based on machine learning*. Submetido à revista ***PLOS Neglected Tropical Diseases***.

1.6.2-Trabalhos em Anais de Eventos

1. SANTOS, Laércio Ives; D'ANGELO, Marcos Flavio Silveira Vasconcelos; MENDES, João Batista & COSTA, Ian D'Angelis. Inteligência de Enxames e Conjuntos Difusos

para monitoramento da saída de idosos do leito**. Anais do LII Simpósio Brasileiro de Pesquisa Operacional**. João Pessoa. Brasil 2020.

1.6.3- Resumos Apresentados

1. GARCEZ, Aline Da Silveira; SANTOS, Laércio Ives; MENDES, João Batista & COSTA, Ian D'Angelis. Predição de Acidente Vascular Cerebral com Árvores de Decisão e Programação Genética**. Anais do XV Fórum de Ensino, Pesquisa, Extensão e Gestão**. Montes Claros, Brasil. 2020.
2. COSTA, Ian D'Angelis; SANTOS, Laércio Ives; MENDES, João Batista & GARCEZ, Aline Da Silveira. Uma abordagem alternativa baseado em Enxames Inteligentes e Conjuntos Difusos para monitoramento e detecção da saída de idosos do leito. **Anais do XV Fórum de Ensino, Pesquisa, Extensão e Gestão**. Montes Claros, Brasil. 2020.

1.7 Estrutura da Tese

Além deste capítulo introdutório, o restante desta tese está organizado como se segue: O capítulo 2 apresenta os dois artigos desenvolvidos e o capítulo 3 faz as considerações finais do trabalho.

2 – Artigos

Artigo 1: SANTOS, Laércio Ives; D'ANGELO, Marcos Flávio Silveira Vasconcelos; COSME, Luciana Balieiro; DE OLIVEIRA, Heveraldo Rodrigues; MENDES, João Batista; EKEL, Petr Ya. ***Swarm intelligence and fuzzy sets for bed exit detection of elderly***. Journal of Intelligent & Fuzzy Systems, v. 39, n. 1, p. 1061-1072, 2020. DOI: 10.3233/JIFS-191971.

Artigo 2: SANTOS, Laércio Ives; CAMARGOS, Murilo Osorio; D'ANGELO, Marcos Flávio Silveira Vasconcelos; MENDES, João Batista; DE MEDEIROS, Egydio Emiliano Camargos, GUIMARÃES, André Luiz Sena; PALHARES, Reinaldo Martínez . *Decision Tree and Artificial Immune Systems for Stroke Prediction in Imbalanced Data.* ***Expert Systems With Applications***. v. 191, p. 116221, 2022. DOI: 10.1016/j.eswa.2021.116221.

Artigo 1:

Journal of Intelligent Fuzzy Systems 1 (2009) 1–5
IOS Press

SWARM INTELLIGENCE AND FUZZY SETS FOR BED EXIT DETECTION OF ELDERLY

Laércio Ives Santos [a,b], Marcos Flávio Silveira Vasnconcelos D'Angelo [c,*], Luciana Balieiro Cosme [b], Heveraldo Rodrigues de Oliveira [c], João Batista Mendes [c] Petr Ya. Ekel [d]

[a] *Graduate Program in Health Sciences, Universitary Hospital, Montes Claros, Brazil*
[b] *Federal Institute of Norte de Minas Gerais, Montes Claros, Brazil*
[c] *Department of Computer Science, UNIMONTES, Av. Rui Braga, sn, Vila Maurict'eia, Montes Claros, Brazil*
[d] *Pontifical Catholic University of Minas Gerais, Graduate Program in Electrical Engineering, Belo Horizonte, Brazil*

Abstract. Falls in the elderly are a public health problem because this population tends to have a longer recovery time and consequently longer hospital beds. Studies show that 84% of falls in hospital rooms occur near the bed, that led to strategies to prevent falls in the elderly population have been studied. In this context, this paper presents a schema for the detection and emission of bed exit alerts in the elderly. This schema uses signals derived from RFID sensors processed by a model based on Intelligent Swarm and Fuzzy Sets. The main contribution of this study is the use of a Membership Windows that reduces the effects of missclassification of other strategies. The proposed work evaluated a data set containing 14 elderly aged between 66 and 86 years divided into two rooms. The results show that the presented approach improves the precision and recall in environments with greater uncertainty of classification.

Keywords: Bed Exit Alarms, Elderly Care, Intelligent Swarm, Fuzzy Sets

1. Introduction

Recently, the population demography has undergone drastic changes, placing elderly population in constant growth [15]. This growth may be associated with a reduction in the birth rate, that is, the number of older people has balanced in relation to the number of non-elderly people or is associated with scientific advances in general. Thus, new discoveries in science, elaboration of new technologies and discoveries of new pharmacological substances and sophisticated equipment allow prolongation of life confronting the pathologies [29]. In this way, the developing of new technologies in order to provide greater well-being for this population is a field of research that has attracted interest from several researchers [1,20,24,28].

The falls recurrent by elderly in domestic or hospital environments are a public health problem. Occurrences of these falls and its consequences may lead to the need for health services which generally implies very high costs [10]. A prevalence of 84% of falls with elderly in hospitals occurs during movement in the bedroom and near bed [12]. In addition, falls in this population result in longer hospital stays compared to younger people [11] and it can result in disorders such as anxiety, depression and even loss of independence. Similarly, caregivers and nurses may also be affected by psychological trauma [23].

Therefore, some studies have focused on the development of strategies for monitoring the bed exit of elderly. This monitoring can provide the caregiver

*Corresponding author. E-mail: marcos.dangelo@unimontes.br

or nurse opportunities for intervention and prevention of possible falls [11,34]. Installation of cameras in the rooms of the elderly allows this monitoring efficiently [4,6,8,30]. A 3D Convolutional Neural Network (CNN) was used to predict motion sequences of healthy adults of ultra low resolution clips of depth images [5]. Its experimental results show that the method is efficient to recognize bed exits and movements anterior to the bed exit, which may allow nurses or caregivers to intervene in a timely manner. However, using video cameras makes the method invasive [21]. In addition, experiments were done only in healthy adults and not in elderly.

A non-invasive approach uses low-cost Radio Frequency Identification (RFID) devices and machine learning techniques [29,31,33,36,37]. Sensors are made up of a small, inexpensive, lightweight RFID tag called Battery Wireless Identification and Sensing Platform -WISP or W^{2}ISP [13] and are used in the garments of elderly patients, as can be seen in Figure 1. This technology provides a low computational cost solution using a single accelerometer per people [31] and a good receptivity on the part of the participants for being light, small and allows simultaneous monitoring of the multiple patients [17] without decreasing patients' privacy, which occurs with the use of video cameras.

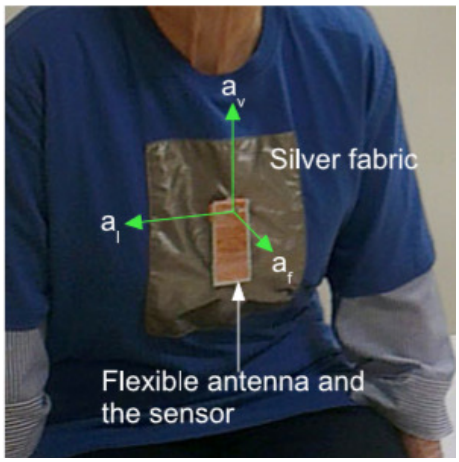


Fig. 1. Elderly using sensor W^{2}ISPens (3g, 18 x 20 x 2 mm), flexible antenna (36 x 85 x 2 mm) and insulating silver fabric (230 x 220 mm). [37]

Limitations of the non-invasive approach are presented: Firstly, it cannot discriminate efficiently the classes that are imbalanced and they are. In additional, seated and standing postures are difficult to discriminate, since in both the shoulders are placed in erect form. Unbalance occurs because the study was performed with elderly who usually spend more time lying down than walking [29]. Secondly is the positioning of some antennas in the room where the experiments were performed [29,32,37]. During the transition of some positions occurred occlusions of some sensors with the body which entails errors of reading and classification and consequently the emission of false alerts. Finally, responsiveness of the bed-exit recognition algorithm can be limited by time the sensor observations are obtained by the antenna and potentially directed to the system [37]. For example, an out-of-bed prediction may occur after the physical transition of a patient because the sensor was not properly energized at the time of the patient's bed exit transition. This is due to the sensor collecting power from the antennas installed in the room and this is a limitation of the sensor used [25]. In [37], the efficiency of the approach depends on the correct postural transition. Thus, reading errors of sensors and or of the classification may compromise the issuance of the alert.

The dificulty in discriminating seated from standing postures, the complexity of dealing with problems with unbalanced classes [29] and the dependence on the correct classification transition sequence [37] will be addressed in this study. Therefore, this paper presents a bed exit monitoring and detection scheme of elderly. This scheme uses signals issued by RFID sensors[29,37] and a prediction system inspired by Particle Swarm Optimization (PSO), denominated New Constructive Particle Swarm Clustering (NcPSC) in this paper, associated with a decision-making model based on Fuzzy Sets [38].

The main contribution of this study is the use of membership values in each class to detect bed exit alerts (section 2.2), reducing dependence on correct postural transition mentioned in [37] and the effects imposed by structural limitations of the sensors and configuration of the experiments [29,32,37]. Another contribution is the use of a classification method based on prototypes, different from other studies. That sort of classification better deals with problems in which classes are arranged alternately and are non-linearly separable.

The efficiency of the proposed scheme is evaluated in a database containing 14 elderly participants aged 66 to 86 years divided into two rooms and the results are compared to ones presented in [29,32,33,37]. Our propose presented better results for room 1 than all other studies. It is important to mention that room 1

has a higher signal loss rate than room 2 due to the way the antennas were installed.

This paper is written in then following manner: similar works are in subsection 1.2, the proposed scheme is described in section 2; section 3 presents the data set used in the experiments and their results; finally, in section 4 the final considerations are made.

1.1. Related works

A real-time data flow tracking strategy is proposed in [36] in which the classification task is performed only in follow-ups whose information is accurate for reducing classification errors. Four classification models, such as Conditional Random Fields (CRF) [18], Support Vector Machine (SVM) [35], Naive Bayes [26] and Random Forest [2] are used to discriminate the activities. The proposed strategy presented good precision and recall values, but it should be emphasized that the experiments were performed in young adults (mean age of 26 years) and not in elderly.

Other studies investigated healthy older adults because usually elderly spend more time lying in bed than performing other activities unlike young people. So [29] uses signals provided by RFID sensors and a method based on CRF for movements recognition considered from bed exit. The strategy presented good results (accuracy and recall values) for of one room, however, in another room its recall value is less than 70% making intervention difficult.

The work proposed in [33] uses a score function in each class (computed as the number of occurrences of the class in fixed time window) and determines as predicted activity the one with the highest value. This mechanism is implemented by Weighted Support Vector Machine (WSVM) classifier [19] for discriminating classes. This study implemented a fixed window with 4.8 s (delay), thus, if intervener is not close to the bed, delay tends to increase reducing the possibility of correct intervention.

In [37], a Sequential Learning Classifier was proposed to analyze bed exit motion. For validation, the authors used data extracted from 14 elderly (66-86 years old) who wore a wearable embodiment of a battery-operated accelerometer RFID sensor loosely attached to their clothes at chest height. The participants undertook a series of activities including bed exit in two room settings.

Finally, [32] presented a CRF-based hierarchical classifier for bed exit detection. The classifier is evaluated using 3 datasets formed by Hospitalized Elderly and Healthy Elderly. Main advantages of the proposed scheme are ability to adapt the trained model to different activities and a shorter training time compared to other models. However, similar to most methods, this approach generates a large amount of false alarms due to signal loss and incorrect readings.

2. Detection and Alerting of Bed Exit Motion

This section presents the proposed approach, illustrated in Figure 2, for issuing bed exit alerts. Initially, labeled instances (containing the class) were used for model learning (training step). In this step, the labeled data are presented to the NcPSC method [27] and it generates a set of labeled particles that maximizes the accuracy of the dataset (subsection 2.1). These particles can be seen as prototypes of groups, that is, a class can be represented by more than one particle . Because the classification process was performed with prototypes and using Euclidean distance as similarity metric, the data were normalized in the interval [0,1], avoiding one dimension from being more important than other. In a next step, the "Computing Membership Function Values" module determines the association values that will compose each window (subsection 2.2). These values are computed using RFID sensor signals and the set of labeled particles found in the training step. These windows are used to determine if a bed exit alarm will be triggered in "Bed Exit Detection and Alerting" module.

2.1. Classification Method

Classification stage used New Constructive Particle Swarm Clustering (NcPSC) method. Firstly, this method was proposed for classifing faults in a drive system of a DC motor [3]. The NcPSC classification mechanism occurs adaptively since algorithm seeks to increase the number of prototypes for increasing the accuracy of the method. However the number of prototypes can not enhance significantly, not allowing to increase the computational cost. In the NcPSC the prototypes are called particles, since it is based on the algorithm Particle Swarm Clustering (PSC) [9], an algorithm for texts grouping adapted from the Particle Swarm Optimization (PSO) method [14].

NcPSC is a rating method inspired by the PSO, and therefore, it implements memory and cooperation concepts for moving particles through the search space. A term of self-organization [16] is added to the particles

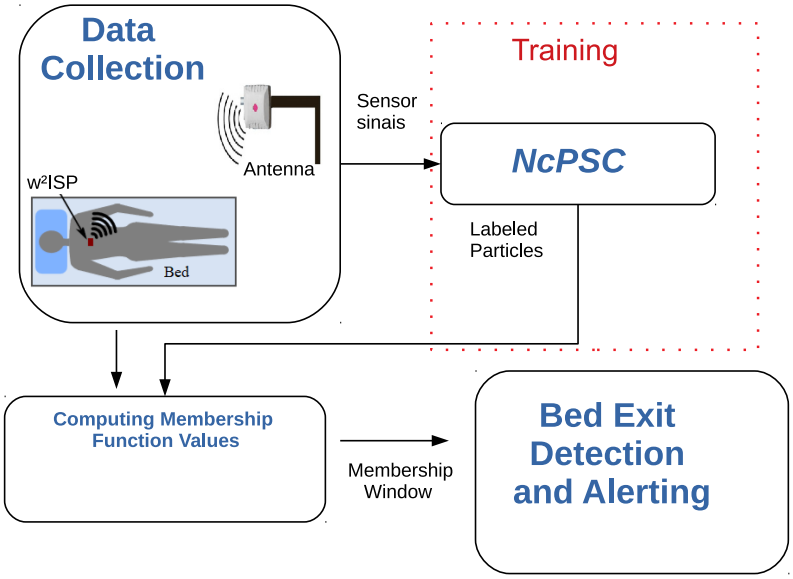


Fig. 2. Classification and Bed Exit Detection scheme flowchart

motion mechanism[9]. Thus, each training instance is presented to the NcPSC and the particle with highest similarity to the instance moves towards it.

Equations 1 and 2, proposed in [9], update, respectively, the speed and position of a particle through the search space. Where: α is the inertia weight used to prevent particles with very high speeds from moving out of the discourse universe. The terms φ_1, φ_2 and φ_3 are vectors of random weights generated in an uniform distribution in the interval [0,1], are used for pondering the weights of memory, cooperation and self-organization, respectively, of each particle; p_w is the winner particle (most similar to the training instance Y_j); and V_w denotes the speed of particle p_w. $Pbest$ and $Gbest$ are, respectively, be the optimal (highest similarity) position found by each particle and by the total swarm in relation to the training instance Y_j.

$$V_w(t+1) = \alpha V_w(t) + \varphi_1(Pbest_{wj}(t) - p_w(t)) + \varphi_2(Gbest_j(t) - p_w(t)) + \varphi_3(Y_j - p_w(t)) \quad (1)$$

$$p_w(t+1) = p_w(t) + V_w(t+1) \quad (2)$$

The NcPSC method is described in algorithm 1. With the following properties: ε_1 and γ_1 are cloning thresholds; ε_2 and γ_2 are movement/stagnation thresholds; $p.CL$ is the concentration level of particle p; $p.HR$ is the hit rate of particle p; η is learning rate; The initial swarm is composed by set of centroids found with the Subtractive Clustering(SC) method [7].

2.1.1. Concentration Level and Hit Rate

The concentration level of a particle p is computed as the ratio between the total of training instances that particle p concentrates (the particle is most similar to the training instances of all particles) and the total of training instances. This value is also used to calculate the hit rate that is obtained by the ratio between the total of training instances belonging to the majority class, among all the classes concentrated by the particle, and the total of training instances associated to p.

The following mechanism is developed by NcPSC for computing the concentration level for all swarm:

- Initialize the concentration level of swarm with zero ($S.CL \leftarrow 0$)
- For each training instance do:
  1. Found the most similar particle (p_w) to training instance;
  2. Update concentration level of p_w: $p_w.CL \leftarrow p_w.CL + 1$;
  3. Associate the current instance to p_w;

Algorithm 1 NcPSC (Training_Instance, ε_1, γ_1, ε_2, γ_2 , ω)

```
S ← InitializeSwarm(Training_Instance);
while stop condition not satisfied do
    for each particle p ∈ S do
        p.CL ← CalculateConcentrationLevel(p);
        p.HR ← CalculateHitRate(p);
    end for
    for ech iten Y_j ∈ Training_Instance do
        Meet p_w ∈ S being p_w the most similar particle to Y_j;
        update Pbest_wj and Gbest_j;
        if p.CL < ε_2 OR p.HR < γ_2 then
            Updates the speed (V_w) of p using (1) and position of p (p_w) with (2)
        end if
    end for
    for each particle p ∈ S do
        if p.CL = 0 then
            ParticleDelete(p);
        end if
        if p.CL > ε_1 AND p.HR < γ_1 then
            CloneParticle(p);
        end if
    end for
    η ← η * 0.95;
end while
```

The routine $CalculateHitRate(S)$ is given:

– For each particle ($p_j \in S$) do:

1. Compute $MC \leftarrow$ majority class from all instances associated to p_j;
2. $N_p \leftarrow$ Total of instances associated to p_j;
3. $\|MC_p\| \leftarrow$ Total of instances from MC in p_j;
4. do $p_j.HR = \frac{\|MC_p\|}{N_p}$

2.1.2. Particle Stagnation

A particle p is stagnated when $p.CL > \varepsilon_2$ AND $p.HR > \gamma_2$. If p has both high hit rate and high concentration level, it will not move in in current iteration. This behaviour reduces chances of particle p explores regions outside the search space [27].

Experiments with artificial data were performed to verify the effect of Particle Stagnation Mechanism (PSM) on problems with unbalanced classes. The arrangement of these data is shown in Figure 3. The dataset is distributed in 3 classes (Class 1, 2 and 3) with 10, 100 and 100 instances, respectively. The NcPSC was applied to artificial data with PSM and without PSM and the results shown in the table 1 (p-value $<$ 0.05 considering the Mann Whitney test [22])). The Results show that the PSM contributes to the improvement of classification in problems with unbalanced classes.

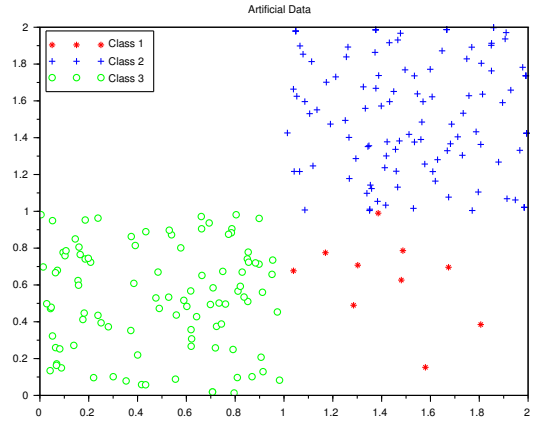


Fig. 3. Disposition of artificial data with unbalanced classes: classes 1, 2, and 3 have, respectively, 10, 100, and 100 instances.

Table 1
Result in Artificial Data

	PSM	**Without PSM**	**p-value**
Overall Result	98.25% ±1.66	93.69% ±2.86	<0.001
Result in Class 1	71.59% ±36.81	14.56 ±30.76%	<0.001

The self-organization mechanism developed by PSC [9] for calculating the velocity of a particle p, causes p to be attracted by the evaluated instance. Therefore regions with high concentration of training instances are more attractive than regions with low concentration of training instances. For example, in Figure 3, while class 2 has 100 instances, class 1 has only 10 instances and a particle located between the two classes will move to class 2 (class with most instances) since class 1 will present the particle only 10 times while class 2 will be presented 100 times. PSM reduces the influence of the amount of instances on particle movement, because when a particle concentrates instances of a minority class but has a good hit rate it will not be attracted to the majority class.

2.1.3. Cloning Mechanism

Figure 4 illustrates the cloning mechanism. When a particle concentrates a reasonable amount of training instances and and it has a low accuracy ($< \gamma_1$) that

particle will be cloned and the new particle undergoes a small change in one of its variables. Figure 4(a) show 2 classes associated to one particle only, so this particle has a low accuracy. Cloning of the particle allows the increase of the accuracy, since part of the data associated to the cloned particle starts to focus on the clone particle. This happens in the illustration showed in Figure 4, part of the instances that were associated to the cloned particle in a) concentrate in the particle clone in b). Note that the instances associated to the two particles are of different classes.

This prototype-based classification model and the cloning mechanism allows the classifier best treat problems in which classes are arranged alternately and are non-linearly separable. Figure 5 ilustrates a problem with 3 classes - C1, C2 and C3. Firstly Figure 5 a) shows the arrangement of the data in the 3 classes. It is easy to note, for example, some class 3 instances are between instances of class 1. To get best results, the NcPSC uses 6 particles for classifing classify the classes instances in micro-group format, as detailed in Figure 5 (b). We assume that classes 1 (sitting in bed) and 3 (standing) are not linearly separable - trunk is erect in both - and so we use NcPSC for the classification task.

2.2. Alerts System

Fuzzy sets are characterized by a membership function able to relate elements of any discourse universe (x) to their respective membership degrees $\mu F(x)$ to the nebulous set F, which can be described as a set of ordered pairs (as in 3). With the modeling in the form of fuzzy sets, x can belong to more than one set simultaneously allowing the model to capture the degree of uncertainty of the problem. The membership function value of an element in the universe of a given Fuzzy set is a real number in the interval [0,1] and represents how true is the assertion that this element belongs to set.

$$F = \{(x, \mu F(x)), x \in X\} \tag{3}$$

In this study, we consider each sensor reading set at time t as x_t with $1 \leq t \leq n$, $x_t \in R^3$ and each class as a fuzzy set represented by FC_i ($1 \leq i \leq 3$, considering that the proposed problem has 3 classes, namely: 1- sitting in bed, 2 - lying in bed and 3- standing). Membership degree of a reading to a given set FC_i is given by (4).

$$\mu FC_i(x_t) = \frac{d(x_t, p_i)}{\sum_{j=1}^{c} d(x_t, p_j)} \tag{4}$$

Satisfying:

$$0 \leq \mu FC_i(x_t) \leq 1, \forall i, t \tag{5}$$

$$\sum_{i=1}^{c} \mu FC_i(x_t) = 1, \forall t \tag{6}$$

In Eq. (4), $d(x_t, p_i)$ denotes the similarity between instance x_t and particle p_i ($p_i \in R^3$) of the C_i class more similar to x_t. As stated earlier, in the proposed scheme, a class can be represented by more than one particle. While constraint (5) deals with fuzzy sets normality, constraint (6) defines that a x_t element must belong to at least one class.

Bed Exit Alert System uses sliding windows composed of $\mu FC_i(x_t)$. As the presented problem has 3 classes, NcPSC also implements 3 sliding windows(W_1, W_2 and W_3) representing, respectively, classes 1, 2 and 3 - all windows with the same size ($|W_1| = |W_2| = |W_3|$). Firstly NcPSC computes the arithmetic mean of each window ($M(W_1)$, $M(W_2)$, and (W_3)) using (7) and a bed exit alert is issued when max($M(W_1)$, $M(W_2)$, $M(W_3)$) = $M(W_3)$ - that is, when the arithmetic mean of sliding window W_3 is greater than arithmetic mean of the other two classes.

$$M(W_i) = \frac{\sum_{j=1}^{|W_i|} W_{ij}}{|W_i|} \tag{7}$$

3. Results and Discussions

3.1. Dataset Performed

To perform the experiments it was used 14 datasets extracted from two rooms: room 1 (09 datasets) and room 2 (05 datasets) [31]. Each data set refers to a participant aged 66-86 years with main characteristics listed in Table 2. For each participant, the readings were performed sequentially and the possible values for each class were observed, wich correspond to the label of the activity of that participant and at that instant of time. These labels can be: 1- sitting in bed, 2

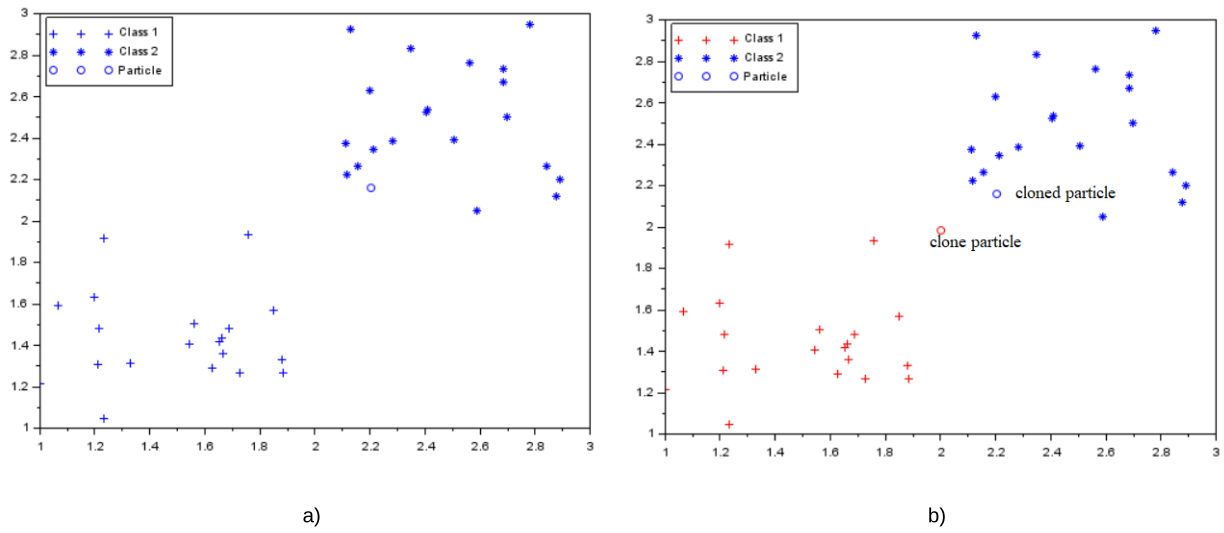


Fig. 4. Cloning Schema: a) Instances of two classes associated to a single particle. b) After Cloning the class instances will focus on the cloned particle and instances of the other class on the clone particle.

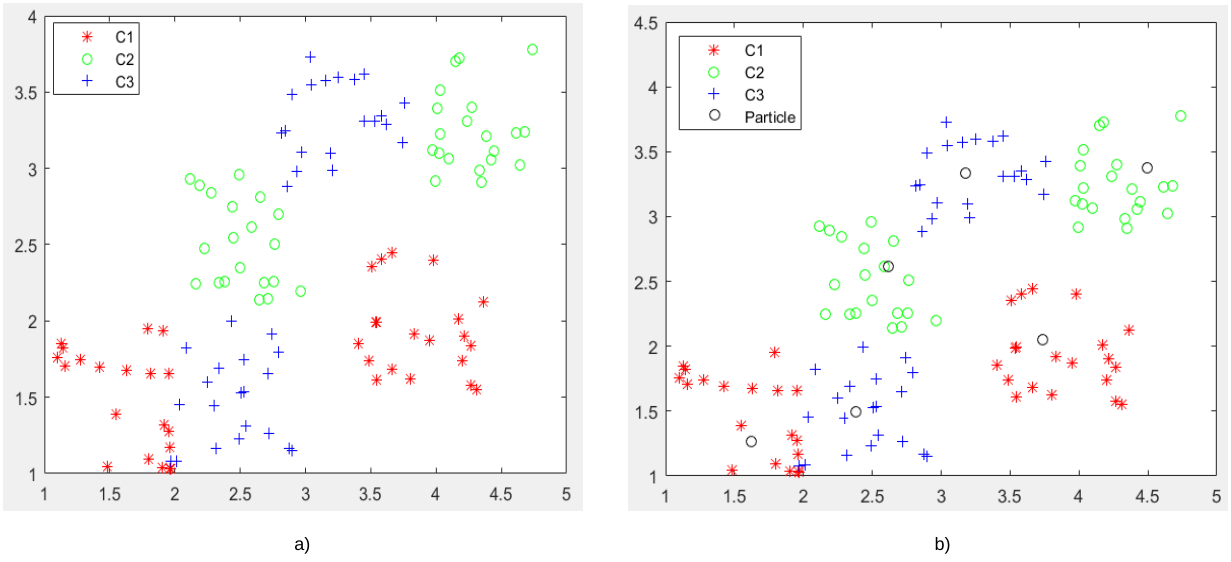


Fig. 5. Illustration of a problem with alternating classes and non-linearly separable: a) Data Set Layout; b) Final arrangement of the set of particles and data

- lying in bed and 3- standing. In this study it is used the time feature only to measure the delay of a bed exit detection. In the data set there are other participants activity, however, in these study, used those who had at least 4 consecutive measurements with label of the class that represents the bed exit motion. This is why a window of size 4 was used.

Table 2
Features Used [31]

Features
Time in seconds
Acceleration reading in G for frontal axis
Acceleration reading in G for vertical axis
Acceleration reading in G for lateral axis
Class

Table 3
RoomSet1 AND RoomSet2 GENERAL INFORMATION

Characteristic	RoomSet1	RoomSet2
Number of antennas	4	3
Number of subjects	9	5
Number of activity sets used	12	6
Number of Number of observations	14030	2455

Participants were randomly divided between the two rooms adapted for elderly monitoring. In room 1 (RoomSet1) there were 4 antennas positioned as: one antenna was placed in a high support at the level of the ceiling facing bed and the others three were putted in vertical supports facing forward. Room 2 (RoomSet2) was configured with 3 antennas: two antennas were positioned in high ceiling-level stands facing the bed and another slightly further away from the bed and facing the chair [31]. Antennas were strategically located to best fit each room arrangements with constraint of ceiling and wall positioning of antennas.

During data collection, each participant wore a garment with the sensor attached over the sternum, as shown in Figure (1) and the participants were informed of the activities that should be performed and each activity was capitalized by the sensor and labeled. Participants performed activities that included:1) lie in bed; 2) sitting in bed; 3) getting out of bed; 4) sitting in the chair; 5) get out of the chair; and 6) going from A to B (A and B represent the bed, chair or door) for 60 to 90 minutes[37].

Two datasets corresponding to each room were evaluated. Table 3 describes differences between both datasets. One of the main challenges of this study is that the data sets used are very unbalanced. For example, in room 1, the reading percentages of classes 1, 2 and 3 are, respectively, 33.47, 61.23, 5.30. In room 2, the percentages are, respectively, 19.20, 74.77 and 6.02. This makes the classification task difficult. Another challenge mentioned in [29] is the difficulty to discriminate the class 1 from class 3 (sitting in bed and standing) since in both the position of the participant's trunk is erect.

3.2. Numerical Results

The efficiency of the presented approach was measured using precision, recall and delay values. Precision measures the rate of True Positives (TP) and in the context of this work, it indicates a correctly triggered alert - we consider TP an alarm triggered at time t when the activity read in $[t-w, t]$ is class 3, w being the size of the window. The precision was calculated in (8) where False Positive (FP) represents the total of incorrectly triggered alerts - the alert was triggered but it should not fire at that moment. The recall was computed in (9) and the False Negative (FN) represents the total of alerts missed - alerts that should be triggered at a given instant but it did not occur. In this study, recall is important because these losses reduce the chances of intervention and possibility of reduction of falls. The delay for triggering an alert is the time elapsed between the instant the alert should be triggered (when ocurs Bed Exit Motion) and the instant the alert is triggered. A high value delay decreases the chances of intervention on the part of the nurse or caregiver and consequently the activity of prevention of falls.

$$P = \frac{TP}{TP + FP} \tag{8}$$

$$R = \frac{TP}{TP + FN} \tag{9}$$

Finally, was used the f-score (Eq. 10) to calculate the amount of FN and FP.

$$f - score = 2 \times \frac{P \times R}{P + R} \tag{10}$$

The experiments were carried out in two distinct ways: First, for each participant the data sets were randomly divided into training sets and test sets (both with same length) as performed in [37]. The precision and recall computed for participants in room 1 and 2 are shown respectively on Table 4 and Table 5. We used 5-fold cross validation in second experiment and its results in terms of precision, recall, f-score and delay values are presented in Table 6. In this configuration, 3 folds were used for training the other two for valida-

tion and testing, respectively.. In the third experiment the participant of each room with largest number of observations in class 3 was used to train the NcPSC. Then, was used the other participants, separately, to test the model. This experiment was performed in this way since in a practical situation, the using a single patient to train the model may be more viable avoiding discomfort to others patients. The results in terms of precision, recall and f-score are presented in Table 7. The results in the two rooms were analysed separately as data were collected differently in each room.

Tables 4 and 5 show the results for the first experiment. As shown in Table 4, our approach produced good precision and recall values for Bed Exit when compared to approach presented in [37]. In the same way, in room 2, present in Table 5, it can see that both are very close. However, these results show a greater heterogeneity of the results of the proposed approach. The overall recall values from rooms 1 and 2 are farther apart in the baseline study. The results of room 1 show that proposed method has lower alarm loss (better recall).

Table 4
Comparison of proposed approach with baseline study proposed in [37] in precision and recall using room 1

	Proposed		**Baseline**	
Participant	**Precision**	**Recall**	**Precision**	**Recall**
1	0.92	0.86	1.00	0.83
2	0.68	0.93	0.60	0.75
3	0.72	0.83	0.50	0.25
4	0.84	0.70	1.00	1.00
5	0.90	0.92	1.00	1.00
6	0.87	0.88	0.86	0.67
7	0.86	0.95	0.83	0.83
8	0.76	0.63	0.50	0.20
9	0.82	0.71	0.83	0.50
Overall	0.82	0.84	0.79	0.67

Table 5
Comparison of proposed approach with baseline study proposed in [37] in precision and recall using room 2

	Proposed		**Baseline**	
Participant	**Precision**	**Recall**	**Precision**	**Recall**
1	0.88	0.50	1.00	1.00
2	0.60	1.00	0.92	0.85
3	0.77	0.72	0.80	1.00
4	1.00	1.00	1.00	0.80
5	0.67	0.83	1.00	0.70
Overall	0.81	0.84	0.85	0.86

A comparison between our methodology and one designed in [29] are presented in Table 6. Firstly, for room 1, recall, precision and delay values produced by NcPSC are better than values presented in [29]. This result is statistically significant (based on an independent t-test) given its p-value (< 0.001)

For room 2, recall, precision and delay values of NcPSC approach are below the approach proposed in [29] in the values of precision, recall, f-score and has a higher delay. But not statistically significant (p.value $>$ 0.05). In both rooms, the proposed approach got good recall. The recall measures the rate of lost alerts. That is, the alert should trigger but this did not occur. In this case, a high recall allows greater possibility of intervention on the part of the caregiver or nurse. This lower loss of alerts associated with a smaller delay can greatly increase the chances of correct intervention and thus, decrease the risk of falls.

Table 6
Comparison of precision, recall,f-score and delay values with in [29]

	Room 1			
Study	**Precision**	**Recall**	**f-score**	**Delay**
Proposed Approach	0.84	0.88	0.86	1.38
dWCRF [29]	0.57	0.64	0.60	3.22
p-value	< 0.001	< 0.001	< 0.001	< 0.001
	Room 2			
	Precision	**Recall**	**f-score**	**Delay**
Proposed Approach	0.72	0.88	0.78	2.80
dWCRF [29]	0.79	0.93	0.85	2.63
p-value	> 0.1	> 0.1	> 0.1	> 0.1

For third experiment, the comparison between our approach and others [29,32,33,37] for room 1 are presented in Table 7. All studies used the same data set in their experiments. Once [36] used young adults in their experiments, we didn't analise it. The results from Table 7 show that only the proposed approach obtained recall and precision values above 80% considering only room 1. Whereas in room 1 there is only one antenna configured facing the bed and in room 2 there are two, the configuration of the antenna in room 1 is more susceptible to signal lossing due to sensor occlusion compared to the configuration of room 2 [29,31]. The results obtained by NcPSC algorithm for room 1 show that this approach presents a lower sensitivity to signal lossing due to occlusion of the sensors.

Efficiency of the approach proposed in [37] depends on correct postural transition detected by a classifica-

Table 7
General comparison in the room 1

Study	precision	recall	f-score	delay
Proposed Approach	0.92	0.93	0.92	1.38
dWCRF [29]	0.57	0.64	0.60	3.22
WSVM [33]	0.67	0.81	0.73	4.8
HCRF [32]	0.42	0.57	0.48	1.7
Sequence Learning [37]	0.79	0.67	0.72	8

tion method and the unavailability of sufficient sensor observations, especially during postural transitions or incorrect classifications, may hinder alert issuances increasing the number of missed alerts (FN). In all other studies[29,32,33], a single misclassification may be sufficient to confuse the system resulting in false alerts. This study implements a window of membership values for each class of the problem, dealing with uncertainties in classification phase. In this context, although a class is predicted incorrectly using the proposed windows mechanism, an alert may be issued correctly, since alerting depends on all window values. This fact is illustrated, for example, in Figure 6, which presents two classes: 1 (sitting in the bed) and 3 (standing). A bed exit alert should occur when the system reports a class 3 instance. For proposed approach this occurs using windows of size 4. In Figure 6, it can be seen that from 4 instances, 2 were classified incorrectly, that is, the system should label instance (x_t) as class 3 but it was labeled as class 1. Still, an alert will be issued when $t = 4$ (window end). The alert is issued since the average value of window 3 is greater than the average of window 1 (Window 3 = [0.46, 0.63, 0.50, 0.40], average value 0.50 and Window 1 = [0.47, 0.30, 0.39, 0.57] , average value 0.43). Although two instances were misclassified (t=1 and t=4), the alarm was issued correctly. Thus, the approach presents is robust regarding unavailable sensor observations and classification errors, as in room 1, for example.

The delay is the mean of difference in time variable between the last and the first observation of a window considered as bed exit movement (containing only activity 3). It has a small value because the windows length is small (4 readings). So, increasing windows length would also increase the waiting time. In some cases there are up to 4 readings in a second, so for these cases, this difference is less than 1 second.

Finally, the main advantages of the proposed approach is discussed:

– Our approach allows greater chance of intervention in possible falls and consequently a greater chance of reducing such falls because it has a lower alert loss (good value of recall) and smaller delay.
– Studies conducted in [29] and [37] use the information from the RFID infrastructure and signals from sensors to perform the prediction and achieve satisfactory performance. NcPSC implements a simple mechanism for bed-exit prediction based on only signals from sensors and obtained similar results.
– Our approach uses labeled prototypes (labeled particles) as micro groups which reduces the misclassification in problems with class overlaping, as shown in Figure 5. In the two rooms there is no clear separation between classes 1 (sitting in bed) and 3 (standing) because in both positions the body is erect [29]. Overlapping classes make data separation difficult and consequently increase classification errors. In this case, NcPSC better discriminated classes by using 1 or more prototypes to represent the same class, as ilustrated in Figure 5 (b).
– When using RFID sensors, the method is inexpensive and allow seniors to move freely and comfortably [37] . This type of sensor requires less maintenance effort because it is not placed directly on the patient's bed [29] and enables simultaneous monitoring of multiple elderly patients which reduces their financial costs. Finally, this approach is non-invasive, which makes it possible to increase acceptance by patients and their families.

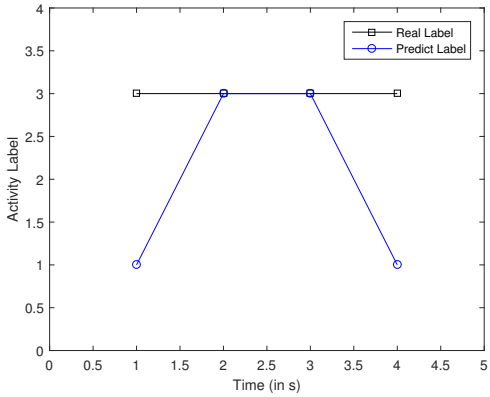


Fig. 6. Illustrative scheme for bed exit detection and alert

4. Conclusions

This work presents an approach for detecting and issuing bed exit alerts. The efficiency of the approach was verified in a problem with elderly between 66 and 86 years. These properly issued alerts can help caregivers and nurses to make interventions in a timely manner reducing falls. The presented approach uses RFID signals and a machine learning technique associated with Fuzzy Sets. Thus, classification errors that occur due to unbalance of the data set in the target class and read errors in the sensors increase the number of missed alarms. Therefore, the NcPSC method performed classification tasks on signals from the RFID sensors in order to minimize classification errors. It was also used a time window composed of membership values in order to minimize the effects of these errors.

The results showed that in room 1, where there is the greatest loss of signals, the present approach obtained best recall and precision values associated to a delay when compared to other approaches of specialized literature. In this sense, the proposed approach can be considered satisfactory in environments with many interferences and consequently many sensor signal losses.

A limitation of our approach is the way it considers a false alarm. Let *w* be window size and *NE3* the total of elements belonging to class 3. A complete window of size *w* has all elements belonging to class 3 ($w = NE3$) while an incomplete window have $0 < NE3 < w$. In room 1 there are 195 complete and 54 incomplete windows while in room 2 there are 35 complete and 49 incomplete windows. As the proposed scheme considers a real movement of bed exit to occur when there a complete window of class 3 (all values are class 3), incomplete windows contribute to issuance of false alerts (FP). For example, an incomplete window (W = [1, 1, 3, 3]) might have Window 1 = [0.39, 0.53, 0.3, 0.2] and Window 3 = [0.3, 0.23, 0.67, 0.59] the average value of Window 3 (0.45) is greater that the average value of Window 1 (0.35), in this context, an alarm will be issued, but as the system considers a real alarm only a full window this emission will be considered an FP.

Another limitation of the present study is the positioning of the antennas which, in many situations, may change the collection of sensor observations, increasing the delay and the number of incomplete windows, consequently increasing the number of False Positives. Future studies may focus on optimizing the efficient positioning of the antennas in order to reduce occlusions and false alert emissions.

In this study, as in others, the number of false alarms (FP) is high, this may make the approach use unviable, because a FP value represents a increase in system cost. In future studies, machine learning methods should investigate additional sources of information to extract robust resources or use robust strategies to discriminate similar activities. In addition, the short duration of experiments doesn't represent the duration of a full day in a hospital. Therefore, long-term trials, including daytime and nighttime, should be considered with a larger sample of hospitalized elderly participants to assess the acceptability of the device after using the sensor for long periods of time.

References

[1] Acharya, U.R., Fujita, H., Oh, S.L., Hagiwara, Y., Tan, J.H., Adam, M., San Tan, R.: Deep convolutional neural network for the automated diagnosis of congestive heart failure using ecg signals. Applied Intelligence 49(1), 16–27 (2019)

[2] Breiman, L.: Random forests. Machine learning 45(1), 5–32 (2001)

[3] Caminhas, W.M., Takahashi, R.H.: Dynamic system failure detection and diagnosis employing sliding mode observers and fuzzy neural networks. In: Proceedings Joint 9th IFSA World Congress and 20th NAFIPS International Conference (Cat. No. 01TH8569), vol. 1, pp. 304–309. IEEE (2001)

[4] Chen, T.X., Hsiao, R.S., Kao, C.H., Liao, W.H., Lin, D.B.: Bed-exit prediction based on convolutional neural networks. In: 2017 International Conference on Applied System Innovation (ICASI), pp. 188–191. IEEE (2017)

[5] Chen, T.X., Hsiao, R.S., Kao, C.H., Lin, D.B., Yang, B.R.: Bed-exit prediction based on 3d convolutional neural network. In: 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), pp. 1185–1188. IEEE (2018)

[6] Chen, T.X., Hsiao, R.S., Kao, C.H., Lin, H.P., Jeng, S.S., Lin, D.B.: Vision-assisted human motion analysis for bed exit prediction model construction. In: 2017 International Conference on Information, Communication and Engineering (ICICE), pp. 542–545. IEEE (2017)

[7] Chiu, S.: Method and software for extracting fuzzy classification rules by subtractive clustering. In: Proceedings of North American Fuzzy Information Processing, pp. 461–465. IEEE (1996)

[8] Chwyl, B., Chung, A.G., Shafiee, M.J., Fu, Y., Wong, A.: Deeppredict: A deep predictive intelligence platform for patient monitoring. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 4309–4312. IEEE (2017)

[9] Cohen, S.C., de Castro, L.N.: Data clustering with particle swarms. In: 2006 IEEE International Conference on Evolutionary Computation, pp. 1792–1798. IEEE (2006)

[10] Heinrich, S., Rapp, K., Rissmann, U., Becker, C., König, H.H.: Cost of falls in old age: a systematic review. Osteoporosis international 21(6), 891–902 (2010)

[11] Hill, K.D., Vu, M., Walsh, W.: Falls in the acute hospital setting-impact on resource utilisation. Australian Health Review 31(3), 471–477 (2007)

[12] Hitcho, E.B., Krauss, M.J., Birge, S., Claiborne Dunagan, W., Fischer, I., Johnson, S., Nast, P.A., Costantinou, E., Fraser, V.J.: Characteristics and circumstances of falls in a hospital setting: a prospective analysis. Journal of general internal medicine 19(7), 732–739 (2004)

[13] Kaufmann, T., Ranasinghe, D.C., Zhou, M., Fumeaux, C.: Wearable quarter-wave folded microstrip antenna for passive uhf rfid applications. International Journal of Antennas and Propagation 2013 (2013)

[14] Kennedy, J.: Particle swarm optimization. Encyclopedia of machine learning pp. 760–766 (2010)

[15] Koh, J.E., Ng, E.Y., Bhandary, S.V., Laude, A., Acharya, U.R.: Automated detection of retinal health using phog and surf features extracted from fundus images. Applied Intelligence pp. 1–15 (2018)

[16] Kohonen, T.: Self-organized formation of topologically correct feature maps. Biological cybernetics 43(1), 59–69 (1982)

[17] Kosse, N.M., Brands, K., Bauer, J.M., Hortobágyi, T., Lamoth, C.J.: Sensor technologies aiming at fall prevention in institutionalized old adults: a synthesis of current knowledge. International journal of medical informatics 82(9), 743–752 (2013)

[18] Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001)

[19] Lin, C.F., Wang, S.D.: Fuzzy support vector machines. IEEE transactions on neural networks 13(2), 464–471 (2002)

[20] Liouane, Z., Lemlouma, T., Roose, P., Weis, F., Messaoud, H.: An improved extreme learning machine model for the prediction of human scenarios in smart homes. Applied Intelligence 48(8), 2017–2030 (2018)

[21] Londei, S.T., Rousseau, J., Ducharme, F., St-Arnaud, A., Meunier, J., Saint-Arnaud, J., Giroux, F.: An intelligent videomonitoring system for fall detection at home: perceptions of elderly people. Journal of telemedicine and telecare 15(8), 383–390 (2009)

[22] Mann, Henry B and Whitney, Donald R, On a test of whether one of two random variables is stochastically larger than the other The annals of mathematical statistics, 50–60 (1947)

[23] Oliver, D.: Prevention of falls in hospital inpatients. agendas for research and practice (2004)

[24] Pandey, Prateek; Litoriya, Ratnesh. An activity vigilance system for elderly based on fuzzy probability transformations. Journal of Intelligent Fuzzy Systems, v. 36, n. 3, p. 2481-2494, 2019.

[25] Ranasinghe, D.C., Sheng, M., Zeadally, S.: Unique radio innovation for the 21st century: building scalable and global rfid networks (2010)

[26] Rish, I., et al.: An empirical study of the naive bayes classifier. In: IJCAI 2001 workshop on empirical methods in artificial intelligence, vol. 3, pp. 41–46 (2001)

[27] Santos, L.I., Palhares, R.M., D'ANGELO, M.F., Mendes, J.B., Veloso, R.R., Ekel, P.Y.: A new scheme for fault detection and classification applied to dc motor. TEMA (São Carlos) 19(2), 327–345 (2018)

[28] Sukor, Abdull et al. A hybrid approach of knowledge-driven and data-driven reasoning for activity recognition in smart homes. Journal of Intelligent Fuzzy Systems, n. Preprint, p. 1-12, 2019.

[29] Shinmoto Torres, R., Visvanathan, R., Hoskins, S., van den Hengel, A., Ranasinghe, D.: Effectiveness of a batteryless and wireless wearable sensor system for identifying bed and chair exits in healthy older people. Sensors 16(4), 546 (2016)

[30] Sokolova, Marina V. et al. A fuzzy model for human fall detection in infrared video. Journal of Intelligent Fuzzy Systems, v. 24, n. 2, p. 215-228, 2013.

[31] Torres R.L.S, Ranasinghe, D.C., Shi, Q., Sample, A.P.: Sensor enabled wearable rfid technology for mitigating the risk of falls near beds. In: 2013 IEEE International Conference on RFID (RFID), pp. 191–198. IEEE (2013)

[32] Torres, R.L.S., Shi, Q., van den Hengel, A., Ranasinghe, D.C.: A hierarchical model for recognizing alarming states in a batteryless sensor alarm intervention for preventing falls in older people. Pervasive and Mobile Computing 40, 1–16 (2017)

[33] Torres, R.L.S., Visvanathan, R., Abbott, D., Hill, K.D., Ranasinghe, D.C.: A battery-less and wireless wearable sensor system for identifying bed and chair exits in a pilot trial in hospitalized older people. PloS one 12(10), e0185,670 (2017)

[34] Vass, C.D., Sahota, O., Drummond, A., Kendrick, D., Gladman, J., Sach, T., Avis, M., Grainge, M.: Refine (reducing falls in in-patient elderly)-a randomised controlled trial. Trials 10(1), 83 (2009)

[35] Wang, L.: Support vector machines: theory and applications, vol. 177. Springer Science & Business Media (2005)

[36] Wickramasinghe, A., Ranasinghe, D.C.: Ambulatory monitoring using passive computational rfid sensors. IEEE Sensors Journal 15(10), 5859–5869 (2015)

[37] Wickramasinghe, A., Ranasinghe, D.C., Fumeaux, C., Hill, K.D., Visvanathan, R.: Sequence learning with passive rfid sensors for real-time bed-egress recognition in older people. IEEE journal of biomedical and health informatics 21(4), 917–929 (2017)

[38] Zadeh, L.A.: Fuzzy sets. Information and control 8(3), 338–353 (1965)

Expert Systems With Applications 191 (2022) 116221

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Decision tree and artificial immune systems for stroke prediction in imbalanced data

Laércio Ives Santos [a,b], Murilo Osorio Camargos [c], Marcos Flávio Silveira Vasconcelos D'Angelo [d,*], João Batista Mendes [d], Egydio Emiliano Camargos de Medeiros [e], André Luiz Sena Guimarães [f], Reinaldo Martínez Palhares [g]

[a] Graduate Program in Health Sciences, UNIMONTES, Montes Claros, Brazil
[b] Federal Institute of Northern Minas Gerais, Montes Claros, Brazil
[c] Graduate Program in Electrical Engineering, Federal University of Minas Gerais, Belo Horizonte, Brazil
[d] Department of Computer Science, UNIMONTES, Av. Rui Braga, sn, Vila Mauricéia, Montes Claros, Brazil
[e] City Hall of Pará de Minas, Pará de Minas, Brazil
[f] Department of Dentistry, UNIMONTES, Montes Claros, Brazil
[g] Department of Electronics Engineering, Federal University of Minas Gerais, Belo Horizonte, Brazil

ARTICLE INFO

Keywords:
Stroke prediction
Artificial immune systems
Decision tree
Genetic programming

ABSTRACT

Although cerebral stroke is a important public worldwide health problem with more than 43 million global cases reported recently, more than 90% of metabolic risk factors are controllable. Therefore, early treatment can take advantage of a fast and low-cost diagnosis to minimize the disease's sequels. The use Machine Learning (ML) techniques can provide an early and low-cost diagnosis. However, the performance of these techniques is reduced in problems of prediction of rare events and with class imbalance. We proposed Machine learning approach to cerebral stroke prediction based on Artificial Immune Systems (AIS) and Decision Trees (DT) induced via Genetic Programming (GP). In general, the approaches for stroke prediction presented in the literature do not allow the development of models considered interpretable; our approach, on the other hand, uses a simplification operator that reduces the complexity of the induced trees to increase their interpretability. We evaluated our approach on a highly imbalanced data set with only 1.89% stroke cases and used AIS combined with One Sided Selection (OSS) to create a new balanced data set. This new data set is used by the GP to evolve a population of DTs, and, at the end of this process, the best tree is used to classify new instances. Two experiments are used to test the proposed approach. In the first experiment, our approach achieved, in terms of sensitivity and specificity, are 70% and 78%, respectively, indicating its competitiveness with the state-of-the-art technique. The second experiment evaluates the proposed simplification mechanism in creating rules that can be interpreted by humans. The proposed approach can effectively increase sensitivity and specificity while maintaining accurate prediction using interpretable models, indicating its potential to be clinically used in stroke diagnosis.

1. Introduction

Despite the scientific advances related to the care of stroke patients in recent years, stroke remains a worldwide public health problem and is among the leading causes of adult death and disabilities (Benjamin et al., 2018; Thrift et al., 2014). There are more than 43 million global cases reported in 2015 (Benjamin et al., 2018) and this amount tends to increase with the growth of the elderly population (Simpkins et al., 2020). In addition, the prevalence of stroke has also increased in the younger population (GBD, 2018). Usually, stroke patients undergo an initial period in the hospital for treatment. In the next stage, they remain an extended period at home for recovering their physical, speech, and cognitive functions (Chen et al., 2019), due to sequels of stroke such as depression and imbalance or loss of physical features (Alghwiri, 2016).

The introduction of early treatment is a way for minimizing sequels of stroke once more than 90% of metabolic risk factors are

* Corresponding author.
E-mail addresses: laercio.ives@gmail.com (L.I. Santos), murilo.camargosf@gmail.com (M.O. Camargos), marcos.dangelo@unimontes.br (M.F.S.V. D'Angelo), joao.mendes@unimontes.br (J.B. Mendes), emiliano.camargos@gmail.com (E.E.C.d. Medeiros), andreluizguimaraes@gmail.com (A.L.S. Guimarães), rpalhares@ufmg.br (R.M. Palhares).

https://doi.org/10.1016/j.eswa.2021.116221
Received 23 December 2020; Received in revised form 24 June 2021; Accepted 10 November 2021
Available online 27 November 2021

controllable (O'Donnell et al., 2016). Clinical exams indicate the stroke diagnostics that can be confirmed by a computed tomography scan, where the gold standard to distinguish the disease's subtypes is the non-contrast computed tomography scan (Wardlaw et al., 2004). However, these image exams can be expensive and inaccessible in regions with difficult access such as rural areas (Leira, Hess, Torner, & Adams, 2008); in such cases, it is possible to use weighted clinical score systems to improve the rapid diagnosis of stroke subtypes (Jin et al., 2016). Other alternatives for diagnosis of stroke include increasing state investments or using Machine Learning (ML) techniques to provide an early and low-cost diagnosis (García-Temza, Risco-Martín, Ayala, Roselló, & Camarasaltas, 2019). ML techniques are interesting because they emulate the human way of thinking and making decisions (El Naqa & Murphy, 2015), analyzes large data sets containing many characteristics in a reasonable time, and can handle complex relationships between data sets, making them more accurate than human specialists in some specific situations (Deo, 2015).

The use of ML techniques for health-related diagnostics tasks meets some challenges; one of them resides in the fact that, compared with healthy subjects, patients with a given disease are generally a small part of the total population. This disproportion in the representation of health and non-healthy subjects is known as the problem of imbalanced data sets, where the class with the highest prevalence is called the majority class, while the rarest class is called the minority class (Haixiang et al., 2017). The challenge in applying ML techniques in handling imbalanced data sets is that they tend to rank all instances in the majority class and none in the minority class, which is generally characterized as the event of most significant interest (Li et al., 2017).

Several papers in the literature used ML techniques for predicting stroke. However, most of them ignore the imbalance of the classes while, in clinical practice, the stroke data set is naturally imbalanced (Liu, Fan, & Wu, 2019). In Colak, Karaman, and Turtay (2015), for example, the authors used Artificial Neural Networks (RNA) or Support Vector Machine (SVM) and a knowledge discovery process to predict stroke. A data set with 167 healthy patients and 130 stroke patients, described by eight clinical variables, was used for training and evaluation of the models. SVMs and Margin-based Censored Regression (MCR) are used as learning algorithms for an automatic feature selection procedure proposed in Khosla et al. (2010) to predict stroke. A comparison of several ML methods that have been applied to predict ischemic stroke is made in Arslan, Colak, and Sarihan (2016). The experiments were performed using a data set with 112 healthy patients and 80 sick patients with SVM presenting best accuracy values.

In Liu et al. (2019), a hybrid approach is described for stroke prediction based on physiological data from a highly imbalanced data set (1.18% of cases of stroke). The hybrid approach is executed in three distinct steps: (i) a data imputation process based on Random Forests (Breiman, 2001) is executed; (ii) the data set is balanced using a methodology that combines Principal components Analysis (PCA) and k-Means clustering methods; (iii) the classification operation is performed by a deep Neural Network with hyperparameters automatically adjusted.

The approach detailed in Liu et al. (2019) presented satisfactory sensitivity and poor specificity. Thus, strategies for improving mainly specificity value without reducing sensitivity value should be investigated. Also, the RNA for prediction is not interpretable, i.e., its results present incomprehensible human terms. In health-related applications, it is interesting to adopt interpretable ML techniques, as they facilitate the problem investigation, generate new insights for solving it, and improve specialists' understanding (Caruana et al., 2015).

The adoption of ML tools in clinical practice requires a careful confirmation of their performance before its use. When the results of a diagnosis test are binary, the discrimination performance is usually measured through sensitivity and specificity (Park & Han, 2018). Sensitivity is defined as the proportion of sick individuals correctly identified with the disease. The specificity, on the other hand, refers to the proportion of non-sick people that are correctly identified without the disease (Park, Choi, & Byeon, 2021).

Therefore, in this work, we propose an alternative approach for stroke prediction on highly imbalanced data sets. The approach, illustrated by Fig. 1, combines both Immune/Neural (D'Angelo et al., 2016) and One-Sided Selection (OSS) (Kubat, Matwin, et al., 1997) techniques to balance the training data and uses Decision Trees (DT) induced by Genetic Programming (GP) (Koza, 1992) for the classification operation. In Fig. 1, $\mathcal{X}_{\text{train}}$ identifies the imbalanced training data, which is summarized in $\mathcal{X}^{+}_{\text{train}}$ by the proposed balancing procedure. The GP algorithm uses $\mathcal{X}^{+}_{\text{train}}$ for evolving a population of DTs. The best decision tree (Decision Tree*), returned by the GP algorithm is used to classify unknown instances.

In this work, we use GP in the induction process instead of traditional strategies such as CART (Breiman, Friedman, Stone, & Olshen, 1984) and C4.5 (Quinlan, 2014) due to their ability for global optimization. These traditional strategies use greedy search in the tree generation process which can lead to sub-optimal solutions. Furthermore, the recursive partitioning in the data set can result in data sets too small for attribute selection in deeper nodes of a tree, overfitting the data (Barros, Basgalupp, De Carvalho, & Freitas, 2011).

In summary, this paper focuses on two main challenges. First, in previous studies using ML for stroke prediction, the data sets used do not suffer from class imbalance. In this situation, the performance of the methods in terms of sensitivity and specificity is heavily compromised. In response to this, we propose a new method for balancing the data set through One Sided Selection and Artificial Immune Systems. This new balancing mechanism is associated with Decision Trees to improve the results of stroke prediction in a highly unbalanced data set when compared to the state-of-the-art in terms of specificity and sensitivity. Second, the algorithms generally applied to stroke prediction problem do not allow the development of models considered interpretable; this type of model is important in health problems because it allows the emergence of new hypotheses related to the problem and their validation by specialists knowledge. Thus, we also present a new simplification operator that reduces the complexity of trees induced by GP increasing interpretability in the resulting models. The remainder of this paper is organized as follows. Section 2 describes the new proposed approach. Section 3 presents the experiments and the results as well as the used data set. Finally, the conclusions are presented in Section 4.

2. Proposed approach

2.1. Immune/neural approach

The Artificial Immune Systems (AIS) are adaptive systems whose development is inspired by theoretical immunology and the known immune functions (Timmis, Hone, Stibor, & Clark, 2008). The AIS constitutes an area in the bio-inspired computation in which abstract components of the immune system are proposed to solve engineering problems (Castro, 2002). Among the immune functions implemented by these components, the basic principles of clonal selection can be used for pattern recognition and optimization problems. The ClonALG proposed in De Castro and Von Zuben (2002) considers different immunological aspects of the clonal selection theory, such as maintenance of a specific memory through selecting and cloning the most stimulated antibodies while pruning the non-stimulated; affinity maturation through hypermutation mechanisms; and selecting clones according to their antigenic affinity. The whole principle is simplified in a way only antibodies compose the immune system, while the antigens constitute the individuals to be recognized.

The balancing mechanism proposed in this paper uses an AIS based on the clonal selection theory to obtain more representative instances within the data sets. In this work, we use a modified version of ClonALG in which the affinity maturation process is aided by a Kohonen neural network (Kohonen, 1990); the result is the immune/neural approach

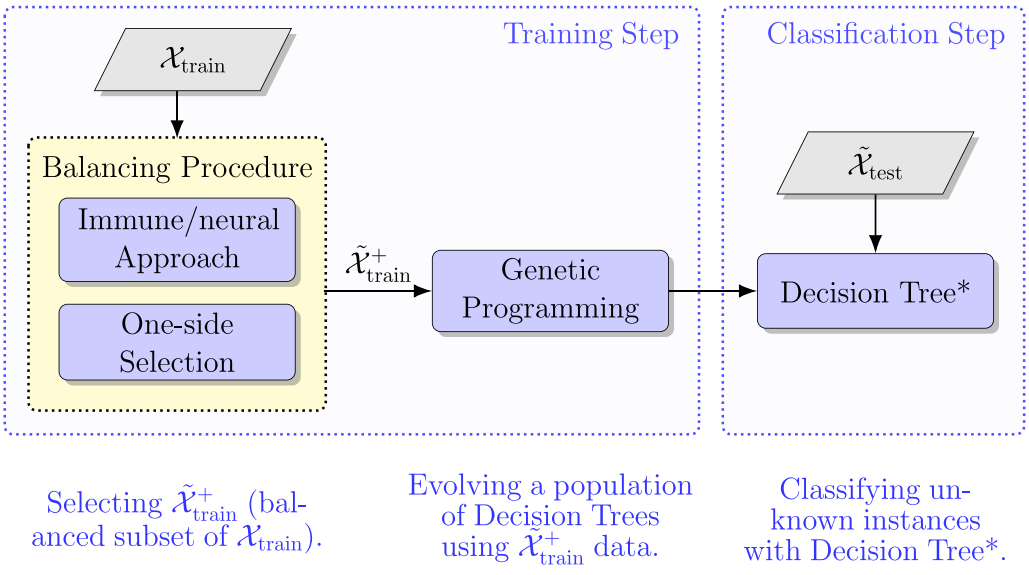


Fig. 1. Overall procedure chart.

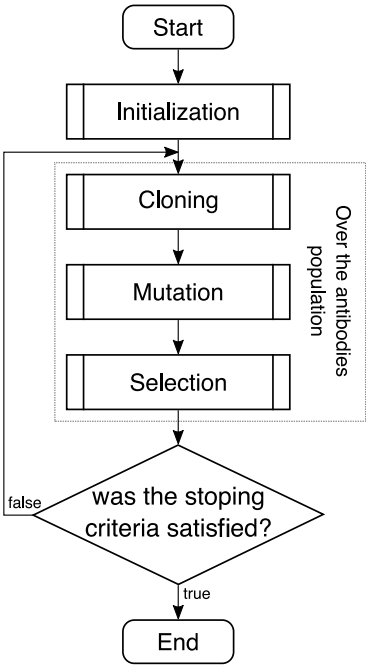


Fig. 2. General flowchart of the used Immune/Neural approach.

described in D'Angelo et al. (2016) and used in Monção et al. (2020) for characterization of salivary gland neoplasms. This structure's training process is made in three steps that are applied to all antibodies in the immune system: cloning, mutation and selection, as shown in Fig. 2. In the cloning step, all antibodies receive two clones that suffer random mutations to ensure diversity in the population. In the hypermutation step, each antibody mutates according to its antigenic affinity; in this step, the weight update procedure of a Kohonen neural network is adopted to increase the antibody affinity to the antigens. In the selection mechanism, the antibodies that recognize the same type of antigen and are close to each other are merged, while antibodies that do not recognize any antigen are pruned. The antibody proximity threshold is computed in each iteration as 25% of the average distance between all antibodies.

The Kohonen network is used in the maturation process of the cloned antibodies. The antibodies are represented by the output units of a Kohonen network in a way their spatial positions are represented by the weight of each unit. The position of the winning unit is adjusted at each iteration with respect to the positions of each antigen (x_i) using the Kohonen network method of weight adjustment. The smallest Euclidean distance between antibodies and antigens defines the winning unit, or winning antibody with respect to that specific antigen. In these terms, the hypermutation mechanism is proportional to the antibodies' affinities since it is directly associated with the distances between antigens and antibodies.

Algorithm 1 Immune/Neural algorithm (IN: $\tilde{\mathcal{X}}^{+}_{\text{train}}$, μ, ψ; OUT: $\mathcal{B}$)

$\mathcal{B} \leftarrow$ Initialize with one centralized antibody
while the stopping criteria is not satisfied **do**
  `// Start cloning section`
  for $\text{ab} \in \mathcal{B}$ **do**
    $\text{ab}^{1}_{\text{tmp}} \leftarrow$ ab clone with random mutation $\mathcal{U}\left(-\frac{\mu}{2}, \frac{\mu}{2}\right)$
    $\text{ab}^{2}_{\text{tmp}} \leftarrow$ ab clone with random mutation $\mathcal{U}\left(-\frac{\mu}{2}, \frac{\mu}{2}\right)$
    $\mathcal{B} \leftarrow \mathcal{B} \cup \{\text{ab}^{1}_{\text{tmp}}, \text{ab}^{2}_{\text{tmp}}\}$
  end for
  `// Start mutation section`
  for $\text{ag} \in \tilde{\mathcal{X}}^{+}_{\text{train}}$ **do**
    $\text{ab}_{\text{win}} \leftarrow$ find the wining antibody in $\mathcal{B}$ with respect to $\tilde{\mathcal{X}}^{+}_{\text{train}}$
    Update ab_{win} position with Kohonen Network updating rule
  end for
  `// Start selection section`
  $\text{affinity} \leftarrow \psi \cdot \text{mean}\left\{\text{norm}(\text{ab}_1, \text{ab}_2) \,|\, (\text{ab}_1, \text{ab}_2) \in \binom{\mathcal{B}}{2}\right\}$
  for $(\text{ab}_1, \text{ab}_2) \in \binom{\mathcal{B}}{2}$ **do**
    if $\text{norm}(\text{ab}_1, \text{ab}_2) < \text{affinity}$ and $\text{ab}_1^{\text{recog AG}} == \text{ab}_2^{\text{recog AG}}$ **then**
      $\text{ab}_{\text{tmp}} \leftarrow$ merges ab_1 with ab_2
      $\mathcal{B} \leftarrow \left(\mathcal{B} \cup \text{ab}_{\text{tmp}}\right) \setminus \{\text{ab}_1, \text{ab}_2\}$
    end if
  end for
end while

Two key parameters are added in the immune/neural approach to allow more liberty in its use for different applications. The random mutation mechanism will be affected by a mutation coefficient $\mu \in [0, 1] \subset \mathbb{R}$ and the proximity threshold is modified by a proximity coefficient

$\psi \in [0,1] \subset \mathbb{R}$. Both parameters will compose the proposed approach's hyperparameters vector and will be optimized in a cross-validation procedure. A short version of the algorithm proposed in D'Angelo et al. (2016) is presented in Algorithm 1.

2.2. One-Sided Selection - OSS

The OSS algorithm is a subsampling method widely used in imbalanced classification problems, i.e., the classification categories are not approximately equally represented (Chawla, 2009). The majority class's most representative instances are selected from a given reference, while all instances of the minority class are preserved. The instance selection from the majority class is made in three steps: (1) selection of a random sample from the majority class; (2) building a data set with all the minority class instances and the instance selected in the first step; (3) the majority class's remaining samples are classified using their nearest neighbor label that belongs to the set constructed in step 2. The instances correctly classified are removed from the data set. The balanced data set is composed of the minority class, the instance selected in step 1, and the instances incorrectly classified in step 3.

2.3. Proposed balancing procedure

Considering an imbalanced data set problem represented by two classes, namely minority and majority classes, the classes set can be defined as $C = \{\min, \text{maj}\}$. The whole data set $\mathcal{X} = \{(\mathbf{x}_1, c_1), \ldots, (\mathbf{x}_n, c_n)\}$ is composed by n pairs $(\mathbf{x}_i, c_i) \in \mathbb{R}^d \times C$ and can be divided into train and test subsets, such that $\mathcal{X} = \mathcal{X}_{\text{train}} \cup \mathcal{X}_{\text{test}}$. The set of antibodies $\mathcal{B} = \{(\mathbf{b}_1, c_1), \ldots, (\mathbf{b}_m, c_m)\}$ obtained through the immune/neural approach is composed by m pairs $(\mathbf{b}_j, c_j) \in \mathbb{R}^d \times C$ and can be divided into minority class antibodies and majority class antibodies, such that $\mathcal{B} = \mathcal{B}_{\min} \cup \mathcal{B}_{\text{maj}}$. The recognition function $\rho : \mathbb{R}^d \times \underline{\mathcal{B}} \to C$ that allows the classification of an antigen $(\mathbf{x}_i, c_i) \in \underline{\mathcal{X}} \subseteq \mathcal{X}$ related to a subset of antibodies $\underline{\mathcal{B}} \subseteq \mathcal{B}$ is given as

$$\rho(\mathbf{x}_i, \underline{\mathcal{B}}) = \arg\max_{c_j} \{\sigma(\mathbf{x}_i, \mathbf{b}_j) \mid (\mathbf{b}_j, c_j) \in \underline{\mathcal{B}}\}, \tag{1}$$

where $\sigma : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is the similarity function, given as

$$\sigma(\mathbf{u}, \mathbf{v}) = \left[\sqrt{(\mathbf{u}-\mathbf{v})^\top (\mathbf{u}-\mathbf{v})}\right]^{-1}. \tag{2}$$

Likewise, the most similar antigen $(\mathbf{x}_i, c_i) \in \underline{\mathcal{X}}$ with respect to a given antibody $\mathbf{b}_j$ can be computed by the function $\beta : \mathbb{R}^d \times \underline{\mathcal{X}} \to \mathbb{R}^d \times C$, in the following way:

$$\beta(\mathbf{b}_j, \underline{\mathcal{X}}) = \arg\max_{(\mathbf{x}_i, c_i)} \{\sigma(\mathbf{x}_i, \mathbf{b}_j) \mid (\mathbf{x}_i, c_i) \in \underline{\mathcal{X}}\}. \tag{3}$$

The proposed balancing strategy is based on a hybrid OSS algorithm aided by the immune/neural approach. A novel scheme for choosing the representative instances in both minority and majority classes is detailed in the following steps:

1. For each antibody in the minority class, obtain the most similar antigen available, such that:

$$\mathcal{S}_{\min} = \{\beta(\mathbf{b}_j, \mathcal{X}_{\text{train}}) \mid (\mathbf{b}_j, k_j) \in \mathcal{B}_{\min}\}. \tag{4}$$

2. For each antibody in the majority class, obtain the most similar antigen available, such that:

$$\mathcal{S}_{\text{maj}} = \{\beta(\mathbf{b}_j, \mathcal{X}_{\text{train}}) \mid (\mathbf{b}_j, k_j) \in \mathcal{B}_{\text{maj}}\}. \tag{5}$$

3. Build a set $\mathcal{S}$ given by the sets found in steps 1 and 2:

$$\mathcal{S} = \mathcal{S}_{\min} \cup \mathcal{S}_{\text{maj}}. \tag{6}$$

4. Classify the other instances in $(\mathbf{x_i}, c_i) \in \mathcal{X}_{\text{train}}$ with the label of its nearest neighbor in $\mathcal{S}$ using the following classification function $\kappa : \mathbb{R}^d \to C$

$$\kappa(\mathbf{x}_i) = \arg\max_{c_s} \{\sigma(\mathbf{x}_i, \mathbf{x}_s) \mid (\mathbf{x}_s, c_s) \in \mathcal{S}\}. \tag{7}$$

5. Build the balanced set composed by the instances incorrectly classified in step 4 and the instances in $\mathcal{S}$, such that:

$$\mathcal{X}^+_{\text{train}} = \{(\mathbf{x}_i, c_i) \mid \kappa(\mathbf{x}_i) \neq c_i, (\mathbf{x}_i, c_i) \in \mathcal{X}_{\text{train}} \backslash \mathcal{S}\} \cup \mathcal{S}. \tag{8}$$

The whole process using a synthetic data set is depicted in Fig. 3. The original data is depicted in Fig. 3(a); the antibodies found by the immune/neural approach are depicted in Fig. 3(b); the most similar antigens from the original data set with respect to the antibodies found are depicted in Fig. 3(c); finally, Fig. 3(d) depicts the balanced data set from (8).

2.4. Decision trees generated by genetic programming

Genetic Programming (GP) is a technique of evolutionary computation that simulates Darwin's principle of natural selection through genetic operators such as reproduction, recombination, and mutation (Banzhaf, 1998). GP systems can represent the candidate solution to a problem in several ways, with trees being quite frequent. In this representation, each individual in the population has ordered ramifications in which the internal nodes are functions while the tree's leaves are the problem's terminals. Each tree is a candidate solution to the problem, and, as in other algorithms in evolutionary computation, they are evaluated with a goodness measure (or fitness) that reflects how good is a solution concerning the others in the same population (Zhao, 2007).

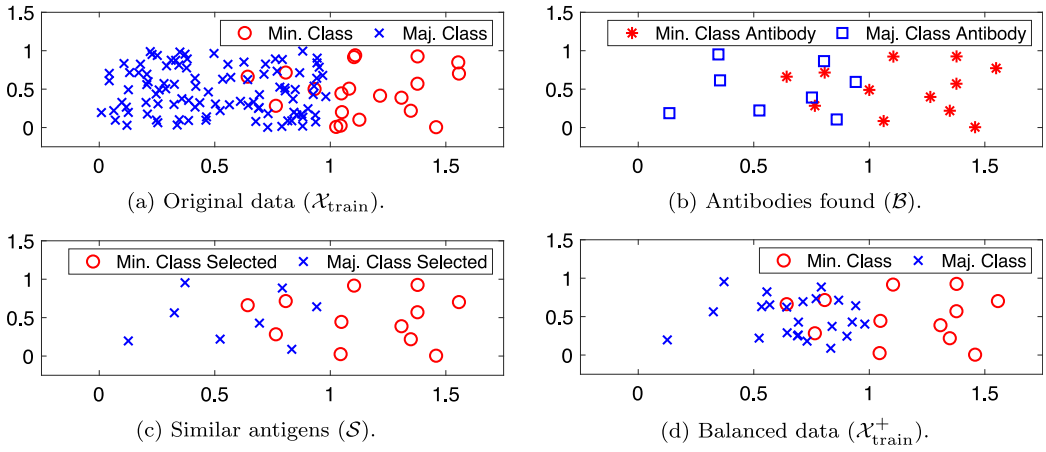
The instance $(\mathbf{x}_i, c_i) \in \mathcal{X}^+_{\text{train}}$, such that $\mathbf{x}_i \in \mathbb{R}^d$, has d attributes that will be preprocessed to assume a finite set of values. The continuous variables will be discretized into N categories using their percentile points while the categorical variables can be re-categorized in the same way when the possible values it can assume are too many, decreasing the final solutions' complexity and overfitting (Saremi & Yaghmaee, 2014). In this paper, we adopt the strategy of limiting the number of percentiles or categories to 4, i.e., $N \in \{2, 3, 4\}$, randomly chosen following a uniform distribution. Therefore, the new training data set $\tilde{\mathcal{X}}^+_{\text{train}}$ is composed by pairs $(\tilde{\mathbf{x}}_i, c_i) \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_d \times C$, where $\mathcal{A}_k = \{a^k_1, \ldots, a^k_{n_k}\}$ is the finite set of n_k values that the kth attribute can assume.

In GP's use to induce decision trees, the internal nodes represent the new training data set's attributes while the leaves represent the classes. The attribute test function $\phi_k : \mathcal{A}_k \to \mathcal{Y} \subseteq \tilde{\mathcal{X}}^+_{\text{train}}$ partitions the new training data set such that each partition contains all instances of $\tilde{\mathcal{X}}^+_{\text{train}}$ where the kth attribute is equal to a given value a^k_j

$$\phi_k(a^k_j) = \{(\tilde{\mathbf{x}}_i, c_i) \mid \tilde{x}_{i,k} = a^k_j, (\tilde{\mathbf{x}}_i, c_i) \in \tilde{\mathcal{X}}^+_{\text{train}}\},$$

where $\tilde{x}_{i,k}$ is the kth attribute of vector $\tilde{\mathbf{x}}_i$. When an instance needs to be evaluated, the function in the tree's root tests the corresponding attribute, and if the argument is a terminal, the decision (classification) for this instance will be returned; otherwise, a new attribute will be evaluated.

The Algorithm 2 shows the pseudo-code for the proposed decision tree induced by GP, where the input parameters are: the new training data set $\tilde{\mathcal{X}}^+_{\text{train}}$, the maximum number of generations $\zeta \in \mathbb{N}^*$, the simplification threshold $\varepsilon \in [0,1] \subset \mathbb{R}$, and the simplification frequency $\tau \in \mathbb{N}^*$, where $\mathbb{N}^*$ is the set of natural numbers without zero; and the output parameter is the best Decision Tree (Decision Tree*). The method Initialization (Line 1) produces the initial population based on the training data set, as described in Section 2.4.1; Fitness (Line 2 and Line 11) measures the fitness of each individual in the population with respect to the training data; the method Recombine (Line 6) constructs a child population using the parent select from the Select method (Line 5); Mutate method (Line 7) inserts variability in the child population, as described in Section 2.4.3; the instruction Remainder (Line 8) guarantees the periodicity of the simplification tests at every τ generations; the Simplify method (Line 9) will prune trees with non-expressive internal nodes, controlling their growth, as described in Section 2.4.5; finally, ε represents the predefined threshold for eliminating sub-trees.



(a) Original data ($\mathcal{X}_{\text{train}}$). (b) Antibodies found ($\mathcal{B}$).

(c) Similar antigens ($\mathcal{S}$). (d) Balanced data ($\mathcal{X}^{+}_{\text{train}}$).

Fig. 3. Balancing procedure example.

Algorithm 2 Genetic Programming(IN:$\tilde{\mathcal{X}}^{+}_{\text{train}}$, ζ, ε, τ; OUT:Decision Tree*)

1: $P \leftarrow$ INITIALIZE($\tilde{\mathcal{X}}^{+}_{\text{train}}$)
2: $F_P \leftarrow$ FITNESS($\tilde{\mathcal{X}}^{+}_{\text{train}}$, P)
3: gen $\leftarrow 1$
4: **while** gen $< \zeta$ **do**
5: $P_S \leftarrow$ SELECT(P, F_P)
6: $P_C \leftarrow$ RECOMBINE(P_S)
7: $P \leftarrow$ MUTATE(P_C)
8: **if** REMAINDER(gen, τ) $= 0$ **then**
9: $P \leftarrow$ SIMPLIFY(P, ε)
10: **end if**
11: $F_P \leftarrow$ FITNESS($\tilde{\mathcal{X}}^{+}_{\text{train}}$, P)
12: gen $\leftarrow$ gen $+ 1$
13: **end while**

2.4.1. Initial population generation

Considering the tuple of all attributes in the data set $\mathbf{A} = (\mathcal{A}_1, \ldots, \mathcal{A}_d)$, let $\mathbf{A}_p$ be the set of all $d!$ permutations of $\mathbf{A}$ where $\mathbf{A}_p(k)$ is the kth permutation. The initial tree population is constructed by choosing a random permutation $\mathbf{A}^* \in \mathbf{A}_p$, such that

$$\mathbf{A}^* = \mathbf{A}_p(u), \quad \text{with } u \sim \mathcal{U}\{1, d!\}, \tag{9}$$

where $\mathcal{U}\{a, b\}$, with $b > a$, is a discrete uniform distribution with support $s \in \{a, a+1, \ldots, b-1, b\}$. The randomly chosen tuple in (9) can be used to construct a tree $\mathcal{T}_{\mathbf{A}^*} = (t_1, \ldots, t_d)$ whose nodes t_n are elements of $\mathbf{A}$. The first element t_1 is the tree's root while the left and right children of element t_n are given by

$$t_{n,\text{left}} = \begin{cases} t_{2n}, & \text{if } 2n \le d \\ \emptyset, & \text{otherwise,} \end{cases} \tag{10a}$$

$$t_{n,\text{right}} = \begin{cases} t_{2n+1}, & \text{if } 2n+1 \le d \\ \emptyset, & \text{otherwise.} \end{cases} \tag{10b}$$

The following trees in the initial population P (Line 1 of Algorithm 2) will be created by moving the elements of $\mathcal{T}_{\mathbf{A}^*}$ to the first position (root node) one by one. After all rotations to create d different trees from $\mathcal{T}_{\mathbf{A}^*}$, a new instance of $\mathbf{A}^*$ will be drawn if the maximum number of trees $\delta \in \mathbb{N}^*$ in the initial population is not reached. This maximum number is empirically defined for each data set and the process to create the initial population is depicted in Fig. 4. The initial population is the set $P = \{p(1), \ldots, p(\delta)\}$ where $p(k)$ is the tree representation of a tuple $\mathbf{A}^* \in \mathbf{A}_p$ whose fitness $f_p(k)$ composes the fitness vector $F_P = \{f_p(1), \ldots, f_p(\delta)\}$.

2.4.2. Selection operator

In order to select which parents will be responsible to create the next generation, a binary tournament is done. The main idea is to run multiple simulations to select parents with greater fitness values and control the selective pressure. Let $\mathcal{I} = \{1, 2, \ldots, \delta\}$ and $\mathcal{I}_p$ is the set of all $\delta!$ permutations of $\mathcal{I}$ where $\mathcal{I}_p(k)$ is the kth permutation. A random permutation $\mathcal{I}^* = \mathcal{I}_p(v)$ is chosen, with $v \sim \mathcal{U}\{1, \delta!\}$, such that

$$\mathcal{I}^* = (i^*_1, \ldots, i^*_\delta), \tag{11}$$

where $i^*_k \in \mathcal{I}$. The selected population is given as

$$P_S = \{p_s(1), \ldots, p_s(\delta_s)\}, \tag{12}$$

where the individuals are chosen by a binary tournament, such that

$$p_s(k) = \begin{cases} p(i^*_{2k-1}), & \text{if } f_p(i^*_{2k-1}) \ge f_p(i^*_{2k}) \\ p(i^*_{2k}), & \text{otherwise.} \end{cases} \tag{13}$$

The number of individuals in P_S is given by

$$\delta_s = \begin{cases} \frac{\delta}{2}, & \text{if } \delta \text{ is even} \\ \frac{\delta-1}{2}, & \text{otherwise.} \end{cases} \tag{14}$$

2.4.3. Variability operators

The recombination procedure takes the parents in the selected population in pairs. Each pair generates two children that will compose P_C, given as

$$P_C = \{p_s(1), \ldots, p_s(\delta_s), p^1_c(1), p^2_c(1), \ldots, p^1_c(\delta_c), p^2_c(\delta_c)\}, \tag{15}$$

where $p_s(k)$ are individuals from the previously selected population and $(p^1_c(k), p^2_c(k))$ are the children generated by crossing $p_s(k)$ and $p_s(r_\delta - k + 1)$ for $1 \le k < \frac{r_\delta + 1}{2}$. The crossover operator chooses a random node on both parents $p_s(k)$ and $p_s(r_\delta - k + 1)$, except their roots; then, the sub-trees whose roots are the random cutting points are exchanged, creating two children: $p^1_c(k)$ and $p^2_c(k)$. An example of the procedure is depicted in Fig. 5; in Fig. 5(a), the first parent $p_s(1) \in P_S$ is chosen and the attribute node $\mathcal{A}_1$ is selected to be a cut point; in Fig. 5(b), the first parent $p_s(\delta_s) \in P_S$ is chosen and the attribute node $\mathcal{A}_9$ is selected to be a cut point; Figs. 5(c) and 5(d) shows the sub-tree exchange between the parents creating two children.

The mutation operator is applied only to previously continuous attributes at a rate of $\eta \in [0, 1] \subset \mathbb{R}$, i.e., the attribute will suffer mutations with probability η. After deciding that a mutation should happen, a previously continuous attribute will be randomly selected to have its discretization limits changed; according to Saremi and Yaghmaee (2018), this mutation mechanism is competitive with others in terms of simplicity and efficiency. For example, suppose a node that represents

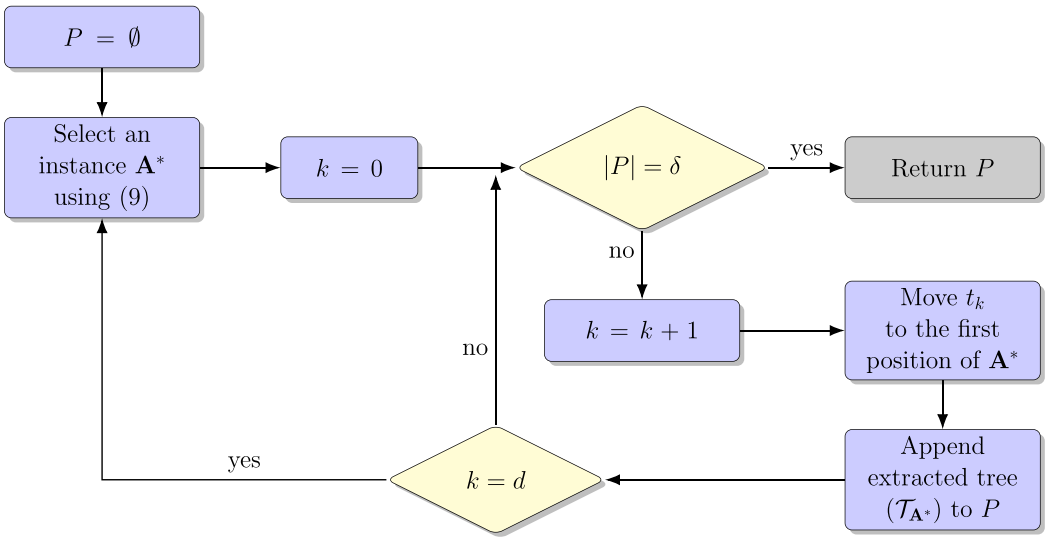
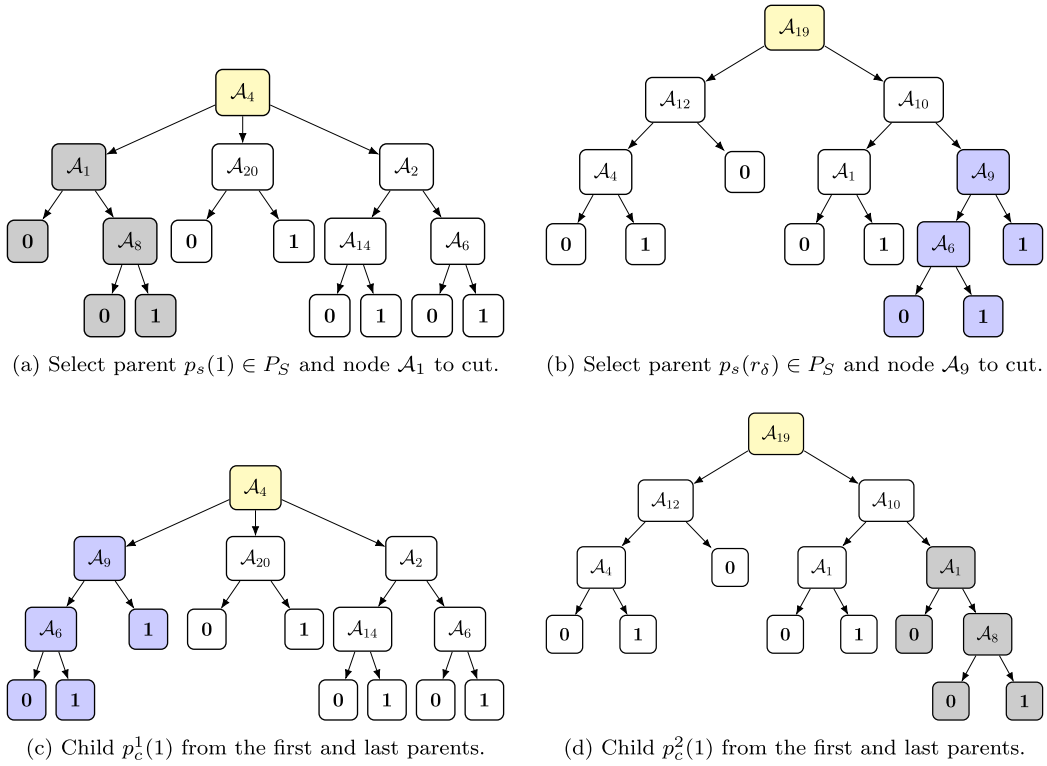


Fig. 4. Initial population generation flowchart.



(a) Select parent $p_s(1) \in P_S$ and node $\mathcal{A}_1$ to cut.

(b) Select parent $p_s(r_\delta) \in P_S$ and node $\mathcal{A}_9$ to cut.

(c) Child $p_c^1(1)$ from the first and last parents.

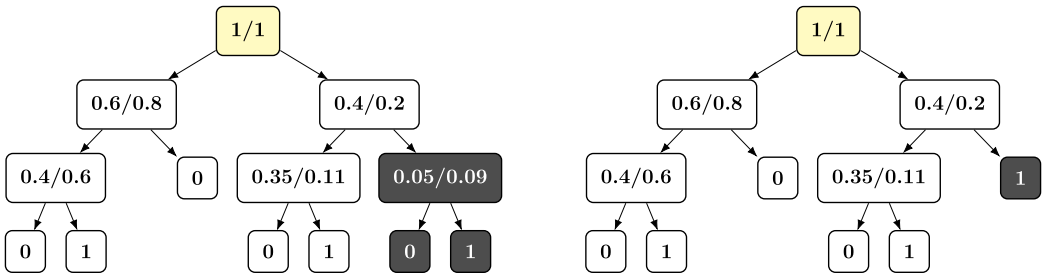
(d) Child $p_c^2(1)$ from the first and last parents.

Fig. 5. Example of the crossover operator behavior applied on a pair of parents $p_s(1), p_s(\delta_s) \in P_S$ generating two children $p_c^1(1)$ and $p_c^2(1)$.

the age attribute in years and the population has individuals whose age lies in the interval $[0, 98]$; a possible discretization of this attribute is to create three categorical ranges, such as $\{[0, 25], [26, 57], [58, 98]\}$. Then a mutation on this discretized attribute could shift the middle interval category in 4 units as, for instance, $\{[0, 25], [26, 61], [62, 98]\}$, creating a new discretization. The new population is assigned to P as the new generation population.

2.4.4. Fitness function

The fitness function evaluates the quality of a given individual with respect to a predefined objective. It is used in the training process to push the population in a given direction aiding in the training process (Devarriya, Gulati, Mansharamani, Sakalle, & Bhardwaj, 2020). In the context of classification tasks, a solution can be evaluated using different metrics such as accuracy, F1-score, G_{mean}, Cohen's kappa, etc (Zhou, Li, & Mitri, 2015). Although accuracy and Choen's kappa are very common evaluating metrics, in imbalanced data sets they can be biased with respect to the majority class, leading to wrongful results (Devarriya et al., 2020). For imbalanced data sets, both F1-score and G-mean metrics minimize the accuracy bias, however, F1-score is preferable when the minority class is more important while G_{mean} is used to maximize the sensitivity of both classes (Al-Badarneh, Habib, Aljarah, & Faris, 2020). Thus, in this paper, the individuals are represented by decision trees and the fitness function is the geometric mean



(a) Selects a sub-tree whose root node represents less than ε% of the training data.

(b) Replaces the sub-tree by a leaf node represented by the most common class in it.

Fig. 6. Example of the simplification operator with $\varepsilon = 0.1$.

between specificity (SPC) and sensibility (SEN), defined as:

$$\text{SPC} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \tag{16a}$$

$$\text{SEN} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{16b}$$

$$G_{\text{mean}} = \sqrt{\text{SPC} \cdot \text{SEN}}, \tag{16c}$$

where TP, FP, TN and FN denote, respectively, true positives, false positives, true negatives and false negatives. This metric is used due to its property of independence to the distribution of examples from different classes, forcing the learning system to produce correct classifications in a significant fraction of the positive examples (Kubat, Holte, & Matwin, 1998).

2.4.5. Simplification operator

The initialization method creates trees with a fixed amount of internal nodes; however, the recombination procedure can make some trees grow boundless over new generations. As the tree size increases, the results are prone to overfitting and can experience a decrease in its interpretability capabilities (DeLisle & Dixon, 2004). Doerr, Lissovoi, and Oliveto (2019) shows that GP systems for evolving simple Boolean functions formed by the conjunction of some variables will require a logarithmic limit on the tree size. In Lissovoi and Oliveto (2019), it is recommended the use of strategies to reduce the tree's growth by using pre-established limiting thresholds. Thus, a threshold (ε) is defined to remove inexpressive sub-trees, defined as the sub-trees whose root node contains less than ε% of training instances from both classes; each node stores the amount of data of each class it represents concerning the entire data set. This procedure is executed periodically with a period τ empirically defined. An example of this procedure considering $\varepsilon = 0.1$ is depicted in Fig. 6, where the gray sub-tree in Fig. 6(a) represents 5% of class 0 and 9% of class 1. Fig. 6(b) shows the replacement of this sub-tree by a leaf node with the most representative class 1.

3. Experiments and results

3.1. Data set

In order to evaluate the proposed approach, the present work uses the same data set evaluated in Liu et al. (2019). The full data set is provided in Liu (2019). The data set is composed of 43,400 instances with ten features, as described in Table 1. In this work, all cases with missing values for at least one feature were removed. The remaining data set is a typical imbalanced data set containing 29,063 instances, with 1.89% of stroke occurrences.

Table 1
Data set description.

Feature	Values	Feature	Values
Gender (gen)	Male/Female	Hypertension (hyp)	Yes/No
Residence type	Urban/Rural	Age	0.08–82
Avg. glucose (glu)	55–291	Heart disease (htd)	Yes/No
Work type (work)	Private/Employed	BMI	10.1–97.6
Married (mar)	Yes/No	Smoking status	*S/F/N

*S/F/N represents Smoked/Formerly/Never.

3.2. Experimental setup

Two experiments are proposed to evaluate the proposed method. In the first experiment, the decision tree induced by the GP algorithm is compared to a state-of-the-art technique in terms of five metrics: specificity (16a), sensitivity (16b), G_{mean} (16c), Area Under the Curve (AUC) derived from a Receiver Operating Characteristic (ROC) analysis and accuracy (ACC), which is defined as:

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}. \tag{17}$$

where TP, FP, TN and FN denote, respectively, true positives, false positives, true negatives and false negatives. The first experiment consists of a 5-fold cross-validation procedure that uses three folds for training, one to adjust the hyperparameters and the last to evaluate the best decision tree obtained. Then, the proposed simplification operator is evaluated in terms of the previous metrics and the average number of internal nodes, referred to as complexity. For these experiments with cross-validation, results are considered statistically significant when the p-value < 0.05 in the comparison, according to the Mann Whitney test (Mann & Whitney, 1947).

In the second experiment, a decision tree obtained is evaluated on a qualitative basis, connecting the current medical knowledge on the subject with the tree generated by data, where 60% of the data set were used for training, 20% to adjust the hyperparameters, and 20% for testing. The hyperparameters vector adjusted on each experiment is given as

$$\boldsymbol{\Theta} = \begin{bmatrix} \mu & \psi & \zeta & \delta & \eta & \tau & \varepsilon \end{bmatrix}^{\top}. \tag{18}$$

3.3. Numerical results and analysis

The proposed balancing procedure performs a sub-sampling in both classes, intending to eliminate less representative instances. This strategy allows the maximization, simultaneously, both sensitivity and specificity. In highly imbalanced data sets, such as the one used in this paper, the proposed mechanism tends to remove more instances in the majority class at the expense of the minority one, as shown in Fig. 7. The average reduction of instances in the majority class is 86%, while in the minority class is 15.5%, considering the five configurations in

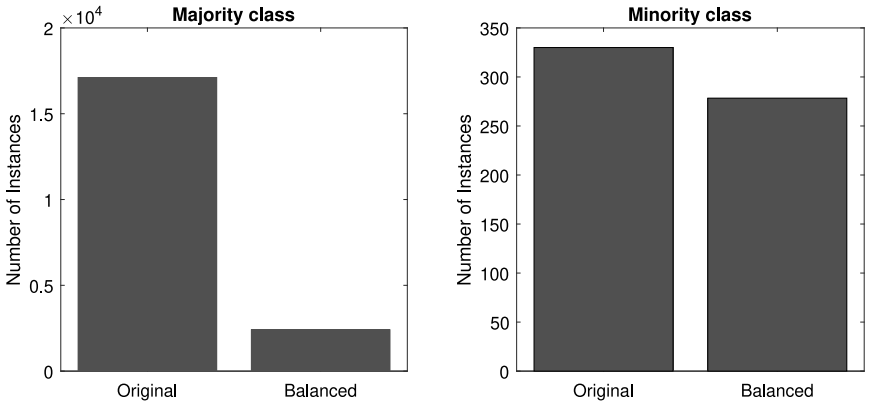


Fig. 7. Instance reduction after balancing.

Table 2
Comparison between the proposed approach and the state-of-the-art technique in terms of four metrics with best values in bold.

Approach	ACC	SPC	SEN	G_{mean}	AUC
Proposed	0.70	**0.70**	**0.78**	**0.74**	**0.74**
AutoHPO Liu et al. (2019)	**0.72**	0.33	0.67	0.47	0.50
p-value	0.713	< 0.002	< 0.002	< 0.002	< 0.002

Table 3
Confusion matrix for proposed approach and state-of-the-art approach.

Approach	Desired output	Predict output	
		Stroke	No stroke
Proposed	Stroke	429	121
	No stroke	8515	19998
AutoHPO	Stroke	368	182
	No stroke	19105	9409

Table 4
Evaluation of the simplification operator in terms of five metrics with best values in bold.

Approach	ACC	SPC	SEN	G_{mean}	AUC	Complexity
With operator	0.70	0.70	**0.78**	**0.74**	0.74	**4.2**
Without operator	**0.76**	**0.76**	0.71	0.73	0.73	17.6
p-value	0.117	0.117	0.14	0.347	0.342	0.009

the cross-validation. Thus, after balancing is done, the minority class represents, on average, 11% of the data set.

The results for the first experiment are compared with the state-of-the-art technique (Liu et al., 2019) in terms of accuracy, sensitivity, specificity, G_{mean} and AUC, as shown in Table 2, where the accuracy was the only metric with no statistical difference to the state-of-the-art (p-value < 0.05). Maximizing sensitivity is of great importance due to the fact that low sensitivity methods can make preventive measures unfeasible, causing irreversible damage to the patients. On the other hand, low specificity can impose a heavy burden to the patients being unfavorable to their health (Liu et al., 2019). According to the results in Table 2, the proposed approach managed to significantly improve sensitivity and specificity in comparison with a state-of-the-art technique; it is essential to emphasize that the proposed method's sensitivity is near 80%, placing it above the second quartile considering the sensitivity from the methods described in a recent systemic review of clinical tools for acute stroke assessment (Antipova, Eadie, Macaden, & Wilson, 2019).

Changing the classification threshold of a model will output multiple pairs of sensitivity and specificity values. The relation between sensitivity and specificity creates a receiver-operating characteristic (ROC) curve, which is an effective method for determining the discrimination performance of the model. The most commonly used summary measure of ROC curves is the area under the ROC curve (AUC) that can take any value between 0 and 1. The closer the AUC is to 1, the better the discrimination performance of the diagnostic test (Park & Han, 2018). The AUC is also widely used to evaluate classifiers in imbalanced data sets. From an statistical point-of-view, the bigger its value, the greater the probability that a randomly chosen diseased person is correctly classified instead of a randomly chosen non-diseased person (Hanley & McNeil, 1982). In Table 2, the AUC of the proposed approach is significantly higher then the AUC of the state-of-the-art approach, implying a greater discriminating power between patients with and without stroke. The low AUC value (0.50) achieved by the state-of-the-art approach is related to its low specificity (0.33).

Another way to show a classifier performance is to use the confusion matrix, in which the principal diagonal indicates the correct classifications (TP and TN). Table 3 shows the confusion matrices of the proposed approach and the state-of-the-art approach. As shown in Table 3, the proposed approach can detect 429 patients with stroke out of 550 patients with the disease; in other words, it detects 61 more patients that can start premature treatment. In short, the proposed approach allows a larger number of interventions in people with some risk of stroke and reduces the false alarm rate, compared to the state-of-the-art approach.

The proposed simplification operator is evaluated in terms of complexity, i.e., the average number of internal nodes in the solution, in Table 4. The simplification operator managed to reduce in almost 80% the number of internal nodes ($p < 0.05$) while maintaining the classification quality concerning sensitivity, specificity, AUC, and G_{mean} ($p > 0.05$). The trees generated when the simplification procedure is active are less complex and more interpretable than trees with boundless growth. In the context of medical practice, the interpretability of a model obtained from ML algorithms has been a topic of interest to establish the trust in these tools, allowing their validation in real environments. Therefore, clinical practitioners and other decision-makers in the health area tend to see the interpretability as a priority in implementing and using these tools (Ahmad, Eckert, & Teredesai, 2018).

In the second experiment, the proposed approach achieved 0.74, 0.74, 0.78 and 0.76 for the metrics accuracy, specificity, sensitivity, and G_{mean}, respectively. The induced tree is composed of three internal nodes, as shown in Fig. 8, where the main reported feature is the Age (the tree's root), followed by Heart Disease (htd) and Avg. Glucose (glu). The age is widely accepted as a strong risk factor for stroke (Gan et al., 2020, 2017; Liu et al., 2019). According to Gan et al. (2020), a person with more the 70 years, in comparison with individuals with 40 to 49 yeas, has approximately 20% more chances of having a stroke. In the induced tree, this knowledge appears in the rule where (Age ≥

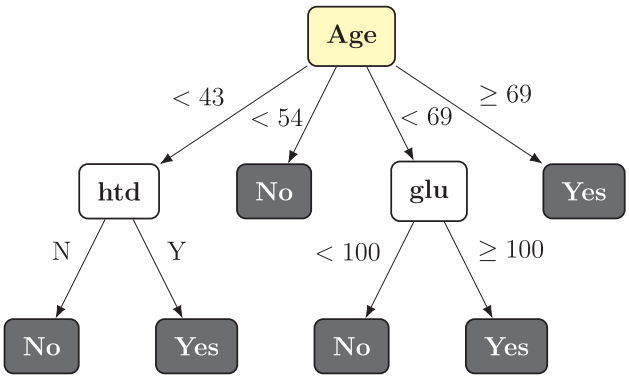


Fig. 8. Tree Induced via GP approach.

69 → Class = yes). Another risk factor associated with stroke, is the presence of Diabetes Mellitus; the presence of this disease can increase in 62% the chances of having a stroke (Gan et al., 2020). This relation is shown in the rule (54 ≤ Age < 69 ∧ glu ≥ 100 → Class = yes), where a high level of glucose is a risk factor for developing diabetes mellitus (Mellitus, 2005). The relationship between blood glucose levels and stoke prediction is also found in the Los Angeles Prehospital Stroke Screen (LAPSS), where blood glucose between 60 and 400 does no exclude true strokes (Kidwell, Starkman, Eckstein, Weems, & Saver, 2000). The occurrence of Heart Disease (htd) as a risk factor in stroke cases was justified in a recent study (Ranganai & Matizirofa, 2020) with private and public hospitals from South Africa where 75.7% of confirmed stroke cases in these hospitals had heart problems and 86.9% had diabetes; this relationship is modeled through the rule (Age < 43 ∧ htd = Y → Class = yes). The presence of glucose and heart diseases as risk factors are important to stroke prediction models because they are controllable factors.

In summary, the proposed balancing strategy eliminates instances considered unrepresentative from both classes; this allows the maximization of sensitivity and specificity, which does not occur in the state-of-the-art technique where the specificity is low. It classifies new instances using a decision tree that can be interpreted by human experts; this interpretability is aided by the complexity reduction achieved through the proposed simplification operator. Using a GP algorithm to induce the creation of decision trees brings two benefits to greedy methods: (i) since GP is a global optimization method, it tends to mitigate the possibility of reach local maxima; (ii) the evolutionary algorithms are more cable of dealing with the complex interactions between attributes and discovering these relations through the concepts of evolution (Barros et al., 2011). It is important to note that the data set used was obtained in a non-invasive way allowing the stroke prediction at a low cost.

4. Conclusion

In this paper, we have presented a novel approach for stroke prediction based on decision trees generated through GP aided by an immune/neural AIS. The proposed approach was evaluated in a highly imbalanced data set composed of sick and non-sick patients' physiological data. The main objective was to present a technique capable of dealing with the imbalance present in the data set while providing a solution that can be interpreted by human specialists. The results have illustrated the achieved improvement in sensitivity and specificity compared to a state-of-the-art model. It also provided an interpretable solution whose complexity management has been done through the proposed simplification operator.

Nevertheless this paper does not address the problem of incomplete data. Unlike the state-of-the-art approach used for comparison, we remove incomplete examples from the data set without any policy to replace missing data with imputed values. Thus, future research directions include studying the proposed approach's robustness to incomplete data sets. Another limitation of the proposed approach is regarding its computational cost; processing times in the order of seconds, minutes, or even days are reported in Espejo, Ventura, and Herrera (2010) for classification tasks using GP. Although high computational costs are common in GP models, due to the individual fitness computation that must be repeatedly evaluated in the evolving process, a parallel version of the proposed algorithm is also an object of future research.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was supported by grants from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil (Grant Number s: 307933/2018-0 and 309909/2019-8), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil and the Fundação de Amparo a Pesquisa do Estado de Minas Gerais (FAPEMIG), Brazil (Grant Number: PPM-00053-17).

All Authors contributed equally to this work.

References

Ahmad, M. A., Eckert, C., & Teredesai, A. (2018). Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics* (pp. 559–560).

Al-Badarneh, I., Habib, M., Aljarah, I., & Faris, H. (2020). Neuro-evolutionary models for imbalanced classification problems. *Journal of King Saud University - Computer and Information Sciences*, http://dx.doi.org/10.1016/j.jksuci.2020.11.005, URL: https://www.sciencedirect.com/science/article/pii/S1319157820305309.

Alghwiri, A. A. (2016). The correlation between depression, balance, and physical functioning post stroke. *Journal of Stroke and Cerebrovascular Diseases*, *25*, 475–479. http://dx.doi.org/10.1016/j.jstrokecerebrovasdis.2015.10.022.

Antipova, D., Eadie, L., Macaden, A., & Wilson, P. (2019). Diagnostic accuracy of clinical tools for assessment of acute stroke: A systematic review. *BMC Emergency Medicine*, *19*, http://dx.doi.org/10.1186/s12873-019-0262-1.

Arslan, A. K., Colak, C., & Sarihan, M. E. (2016). Different medical data mining approaches based prediction of ischemic stroke. *Computer Methods and Programs in Biomedicine*, *130*, 87–92. http://dx.doi.org/10.1016/j.cmpb.2016.03.022.

Banzhaf, W. (1998). *Genetic programming: an introduction on the automatic evolution of computer programs and its applications*. San Francisco, Calif. Heidelburg: Morgan Kaufmann Publishers Dpunkt-verlag.

Barros, R. C., Basgalupp, M. P., De Carvalho, A. C., & Freitas, A. A. (2011). A survey of evolutionary algorithms for decision-tree induction. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*, 291–312. http://dx.doi.org/10.1109/TSMCC.2011.2157494.

Benjamin, E. J., Virani, S. S., Callaway, C. W., Chamberlain, A. M., Chang, A. R., Cheng, S., et al. (2018). Heart disease and stroke statistics—2018 update: A report from the American heart association. *Circulation*, http://dx.doi.org/10.1161/CIR.0000000000000558.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. http://dx.doi.org/10.1023/A:1010933404324.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC Press.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30 day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1721–1730). http://dx.doi.org/10.1145/2783258.2788613.

Castro, L. (2002). *Artificial immune systems: A new computational intelligence approach*. London New York: Springer.

Chawla, N. V. (2009). Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook* (pp. 875–886). Springer US, http://dx.doi.org/10.1007/978-0-387-09823-445.

Chen, Y., Abel, K. T., Janecek, J. T., Chen, Y., Zheng, K., & Cramer, S. C. (2019). Home-based technologies for stroke rehabilitation: A systematic review. *International Journal of Medical Informatics*, *123*, 11–22. http://dx.doi.org/10.1016/j.ijmedinf.2018.12.001.

Colak, C., Karaman, E., & Turtay, M. G. (2015). Application of knowledge discovery process on the prediction of stroke. *Computer Methods and Programs in Biomedicine*, *119*, 181–185. http://dx.doi.org/10.1016/j.cmpb.2015.03.002.

D'Angelo, M. F., Palhares, R. M., Camargos Filho, M. C., Maia, R. D., Mendes, J. B., & Ekel, P. Y. (2016). A new fault classification approach applied to tennessee eastman benchmark process. *Applied Soft Computing*, *49*, 676–686. http://dx.doi.org/10.1016/j.asoc.2016.08.040.

De Castro, L. N., & Von Zuben, F. J. (2002). Learning and optimization using the clonal selection principle. *IEEE Transactions on Evolutionary Computation*, *6*, 239–251. http://dx.doi.org/10.1109/TEVC.2002.1011539.

DeLisle, R. K., & Dixon, S. L. (2004). Induction of decision trees via evolutionary programming. *Journal of Chemical Information and Computer Sciences*, *44*, 862–870. http://dx.doi.org/10.1021/ci034188s.

Deo, R. C. (2015). Machine learning in medicine. *Circulation*, *132*, 1920–1930. http://dx.doi.org/10.1161/CIRCULATIONAHA.115.001593.

Devarriya, D., Gulati, C., Mansharamani, V., Sakalle, A., & Bhardwaj, A. (2020). Unbalanced breast cancer data classification using novel fitness functions in genetic programming. *Expert Systems with Applications*, *140*, Article 112866. http://dx.doi.org/10.1016/j.eswa.2019.112866.

Doerr, B., Lissovoi, A., & Oliveto, P. S. (2019). Evolving boolean functions with conjunctions and disjunctions via genetic programming. In *Proceedings of the genetic and evolutionary computation conference* (pp. 1003–1011). http://dx.doi.org/10.1145/3321707.3321851.

El Naqa, I., & Murphy, M. J. (2015). What is machine learning? In *Machine learning in radiation oncology* (pp. 3–11). Springer, http://dx.doi.org/10.1007/978-3-319-18305-31.

Espejo, P. G., Ventura, S., & Herrera, F. (2010). A survey on the application of genetic programming to classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*, 121–144. http://dx.doi.org/10.1109/TSMCC.2009.2033566.

Gan, Y., Jiang, H., Room, R., Zhan, Y., Li, L., Lu, K., et al. (2020). Prevalence and risk factors associated with stroke in China: A nationwide survey of 726,451 adults. *European Journal of Preventive Cardiology*, http://dx.doi.org/10.1177/2047487320902324.

Gan, Y., Wu, J., Zhang, S., Li, L., Yin, X., Gong, Y., et al. (2017). Prevalence and risk factors associated with stroke in middle-aged and older Chinese: A community-based cross-sectional study. *Scientific Reports*, *7*, 1–7. http://dx.doi.org/10.1038/s41598-017-09849-z.

García-Temza, L., Risco-Martín, J. L., Ayala, J. L., Roselló, G. R., & Camarasaltas, J. M. (2019). Comparison of different machine learning approaches to model stroke subtype classification and risk prediction. In *2019 spring simulation conference* (pp. 1–10). IEEE, http://dx.doi.org/10.23919/SpringSim.2019.8732846.

GBD (2018). Global, regional, and country-specific lifetime risks of stroke, 1990 and 2016. *New England Journal of Medicine*, *379*, 2429–2437. http://dx.doi.org/10.1056/NEJMoa1804492.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, *73*, 220–239. http://dx.doi.org/10.1016/j.eswa.2016.12.035.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve.. *Radiology*, *143*, 29–36. http://dx.doi.org/10.1148/radiology.143.1.7063747.

Jin, H.-Q., Wang, J.-C., Sun, Y.-A., Lyu, P., Cui, W., Liu, Y.-Y., et al. (2016). Prehospital identification of stroke subtypes in Chinese rural areas. *Chinese Medical Journal*, *129*, 1041–1046. http://dx.doi.org/10.4103/0366-6999.180521.

Khosla, A., Cao, Y., Lin, C. C. -Y., Chiu, H. -K., Hu, J., & Lee, H. (2010). An integrated machine learning approach to stroke prediction. In *Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 183–192). http://dx.doi.org/10.1145/1835804.1835830.

Kidwell, C. S., Starkman, S., Eckstein, M., Weems, K., & Saver, J. L. (2000). Identifying stroke in the field. *Stroke*, *31*, 71–76. http://dx.doi.org/10.1161/01.str.31.1.71.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*, 1464–1480. http://dx.doi.org/10.1109/5.58325.

Koza, J. R. (1992). *Genetic programming: on the programming of computers by means of natural selection, volume 1*. MIT Press.

Kubat, M., Holte, R. C., & Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, *30*, 195–215. http://dx.doi.org/10.1023/a:1007452223027.

Kubat, M., Matwin, S., et al. (1997). Addressing the curse of imbalanced training sets: One-sided selection. *97*, In *International conference on machine learning, volume 97* (pp. 179–186). Morgan Kaufmann.

Leira, E. C., Hess, D. C., Torner, J. C., & Adams, H. P. (2008). Rural-urban differences in acute stroke management practices: A modifiable disparity. *Archives of Neurology*, *65*, 887–891. http://dx.doi.org/10.1001/archneur.65.7.887.

Li, J., Liu, L. -s., Fong, S., Wong, R. K., Mohammed, S., Fiaidhi, J., et al. (2017). Adaptive swarm balancing algorithms for rare-event prediction in imbalanced healthcare data. *PLoS One*, *12*, http://dx.doi.org/10.1371/journal.pone.0180830.

Lissovoi, A., & Oliveto, P. S. (2019). On the time and space complexity of genetic programming for evolving Boolean conjunctions. *Journal of Artificial Intelligence Research*, *66*, 655–689. http://dx.doi.org/10.1613/jair.1.11821.

Liu, T. (2019). Data for: A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical-datasets. Mendeley, http://dx.doi.org/10.17632/X8YGRW87JW.1, URL: https://data.mendeley.com/datasets/x8ygrw87jw/1.

Liu, T., Fan, W., & Wu, C. (2019). A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. *Artificial Intelligence in Medicine*, *101*, Article 101723. http://dx.doi.org/10.1016/j.artmed.2019.101723.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 50–60. http://dx.doi.org/10.1214/aoms/1177730491.

Mellitus, D. (2005). Diagnosis and classification of diabetes mellitus. *Diabetes Care*, *28*, S5–S10. http://dx.doi.org/10.2337/diacare.27.2007.s5.

Monção, C. R. L., Santos, E. M., Prates, T. S., de Paula, A. M. B., Cardoso, C. M., Farias, L. C., et al. (2020). Immune/neural approach to characterize salivary gland neoplasms (SGN). *Applied Soft Computing*, *88*, Article 105877. http://dx.doi.org/10.1016/j.asoc.2019.105877.

O'Donnell, M. J., Chin, S. L., Rangarajan, S., Xavier, D., Liu, L., Zhang, H., et al. (2016). Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): A case-control study. *The Lancet*, *388*, 761–775. http://dx.doi.org/10.1016/S0140-6736(16)30506-2.

Park, S. H., Choi, J., & Byeon, J.-S. (2021). Key principles of clinical validation, device approval, and insurance coverage decisions of artificial intelligence. *Korean Journal of Radiology*, *22*, 442. http://dx.doi.org/10.3348/kjr.2021.0048.

Park, S. H., & Han, K. (2018). Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology*, *286*, 800–809. http://dx.doi.org/10.1148/radiol.2017171920.

Quinlan, J. R. (2014). *C4. 5: Programs for machine learning*. Elsevier.

Ranganai, E., & Matizirofa, L. (2020). An analysis of recent stroke cases in South Africa: Trend, seasonality and predictors. *South African Medical Journal*, *110*, 92. http://dx.doi.org/10.7196/samj.2020.v110i2.013891.

Saremi, M., & Yaghmaee, F. (2014). Evolutionary decision tree induction with multi-interval discretization. In *2014 Iranian conference on intelligent systems* (pp. 1–6). IEEE, http://dx.doi.org/10.1109/IranianCIS.2014.6802543.

Saremi, M., & Yaghmaee, F. (2018). Improving evolutionary decision tree induction with multi-interval discretization. *Computational Intelligence*, *34*, 495–514. http://dx.doi.org/10.1111/coin.12153.

Simpkins, A. N., Janowski, M., Oz, H. S., Roberts, J., Bix, G., Doré, S., et al. (2020). Biomarker application for precision medicine in stroke. *Translational Stroke Research*, *11*, 615–627. http://dx.doi.org/10.1007/s12975-019-00762-3.

Thrift, A. G., Cadilhac, D. A., Thayabaranathan, T., Howard, G., Howard, V. J., Rothwell, P. M., et al. (2014). Global stroke statistics. *International Journal of Stroke*, *9*, 6–18. http://dx.doi.org/10.1111/ijs.12245.

Timmis, J., Hone, A., Stibor, T., & Clark, E. (2008). Theoretical advances in artificial immune systems. *Theoretical Computer Science*, *403*, 11–32. http://dx.doi.org/10.1016/j.tcs.2008.02.011.

Wardlaw, J. M., Seymour, J., Cairns, J., Keir, S., Lewis, S., & Sandercock, P. (2004). Immediate computed tomography scanning of acute stroke is cost-effective and improves quality of life. *Stroke*, *35*, 2477–2483. http://dx.doi.org/10.1161/01.str.0000143453.78005.44.

Zhao, H. (2007). A multi-objective genetic programming approach to developing Pareto optimal decision trees. *Decision Support Systems*, *43*, 809–826. http://dx.doi.org/10.1016/j.dss.2006.12.011.

Zhou, J., Li, X., & Mitri, H. S. (2015). Comparative performance of six supervised learning methods for the development of models of hard rock pillar stability prediction. *Natural Hazards*, *79*, 291–316. http://dx.doi.org/10.1007/s11069-015-1842-3.

3-Considerações Finais

A utilização de tecnologias baseadas em Aprendizado de Máquina na área de saúde tem sido uma estratégia interessante adotada por pesquisadores como resposta a diversos problemas. Diante disso, alguns desafios ainda persistem, os mais comuns são: incertezas de classificação, desbalanceamento de classe, busca por modelos que possam ser mais interpretáveis por especialistas humanos e ausência de dados. Nesta tese abordamos 3 desses importantes desafios ao longo de 2 artigos.

Com relação as incertezas de classificação, propomos uma abordagem que utiliza janelas de pertinência (Conjuntos Difusos) para tentar reduzir essas incertezas inerentes ao problema de Detecção da Saída de Idosos do Leito (artigo 1). No problema abordado, as incertezas são causadas, principalmente, pela dificuldade dos algoritmos de classificação em discriminar as atividades “sentado na cama” e “em pé” e, ainda, pela oclusão dos sensores quando o corpo do participante ficava entre o dispositivo preso as roupas e a antena. É importante dizer que os dois problemas poderiam ser minimizados com um melhor posicionamento das antenas nas salas e/ou com a utilização de informações adicionais. Entretanto, os modelos foram treinados com dados disponibilizados publicamente e a concepção de um ambiente experimental para coletar novos dados seria inviável, devido ao momento de pandemia vivido durante a realização da pesquisa,

Ainda na Detecção de Saída de Idosos do Leito, outro desafio abordado foi o desbalanceamento de classes. Para isso, utilizamos um mecanismo de estagnação de partículas, inserido no algoritmo de classificação. Esse mecanismo diminui significativamente o efeito da classe majoritária sobre a movimentação das partículas e o efeito do desbalanceamento na classificação. O desbalanceamento de classes também foi tratado na predição de Acidente Vascular Cerebral (AVC) (artigo 2) por uma abordagem baseada em um Sistema Imunológico Artificial (SIA). A estratégia empregada no artigo 2 demostrou ser interessante pois seleciona instâncias de treinamento mais representativas nas duas classes, e consequentemente, reduz a perda de informações relevantes permitindo maximizar tanto a sensibilidade quanto a especificidade.

Ao lidar com a predição de AVC, optamos por utilizar um modelo construído a partir de uma Árvore de Decisão por que esses modelos, a depender da quantidade de nós internos, são considerados de fácil compreensão e interpretação por especialistas humanos (quando a quantidade de nós internos é grande a interpretabilidade do modelo diminui). Utilizamos um algoritmo evolutivo baseado em Programação Genética (PG) na indução do modelo. A PG

tende a induzir árvores com mais nós internos e mais complexas. Por esse motivo, inserimos um mecanismo de simplificação de árvores para reduzir sua complexidade, melhorar a capacidade de generalização e aumentar a interpretabilidade do modelo. É importante esclarecer que as estratégias empregadas anteriormente nesse conjunto de dados, utilizaram abordagens que dificultam a interpretabilidade por serem modelos do tipo caixa-preta. Mesmo com a utilização de explicadores externos, em modelos do tipo caixa-preta, não há garantias de modelos integralmente interpretáveis.

Como principais pontos fortes desse trabalho, destacamos os avanços no tratamento dos problemas abordados. Esses avanços permitem almejar (em um futuro próximo) a criação de ferramentas inovadoras derivadas das duas abordagens propostas, e com essas ferramentas melhorar a qualidade dos serviços prestados a população idosa, em geral, e melhorar a tomada de decisão a respeito de intervenções em pacientes com risco de desenvolver AVC. O desenvolvimento dessas ferramentas pode, ainda, contribuir para a geração de novos empregos. Os resultados obtidos demonstraram que ainda há espaço para melhorias e trabalhos futuros, tais como:

- Na abordagem apresentada no artigo 2, optamos por excluir instâncias com ao menos uma variável de valor ausente e não pela imputação. Portanto, mais experimentos podem ser realizados para mensurar a assertividade da abordagem na presença de valores ausentes ou quando algum método de imputação é utilizado.
- O desbalanceamento de classes segue com questões ainda em aberto. Por exemplo, nenhuma das abordagens propostas nesse trabalho foi projetada para lidar com o problema quando há sobreposição das classes. Além disso, a abordagem baseada em SIA não foi avaliada em problemas com mais de duas classes. Entretanto, a adaptação da abordagem para essa categoria de problemas pode ser realizada em trabalhos futuros.
- As duas abordagens foram avaliadas em conjuntos de dados disponibilizado publicamente, ou seja, não foi feito uma estudo prévio de quais variáveis preditoras poderiam ser importantes para melhorar a assertividade das abordagens apresentadas e isso pode ser melhor explorado.
- As incertezas inerentes ao problema de Monitoramento da Saída de Idosos do Leito poderiam ser minimizadas com a utilização de mais informações e um melhor posicionamento das antenas nas salas. E isso também pode ser melhor explorado em trabalhos futuros. Ademais, o desenvolvimento integral da ferramenta de

monitoramento a partir de dispositivos *RFID* e algoritmos de Aprendizado de Máquina também faz parte dos objetivos futuros.

Finalmente, esperamos que esta tese possa ajudar na consolidação do uso de ferramentas baseadas em Aprendizado de Máquina na tomada de decisão na área da saúde e fornecer um conhecimento sólido necessário para pesquisas e desenvolvimentos futuros.

Referências

ABDEL-ZAHER, Ahmed M.; ELDEIB, Ayman M. Breast cancer classification using deep belief networks. Expert Systems with Applications, v. 46, p. 139-144, 2016.

AHMAD, Muhammad Aurangzeb; ECKERT, Carly; TEREDESAI, Ankur. Interpretable machine learning in healthcare. In: Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics. 2018. p. 559-560.

AHMADI, Hossein et al. Diseases diagnosis using fuzzy logic methods: A systematic and meta-analysis review. Computer Methods and Programs in Biomedicine, v. 161, p. 145-172, 2018.

AMRANE, Meriem et al. Breast cancer classification using machine learning. In: 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT). IEEE, 2018. p. 1-4.

ARRUDA, Natália Martins; MAIA, Alexandre Gori; ALVES, Luciana Correia. Desigualdade no acesso à saúde entre as áreas urbanas e rurais do Brasil: uma decomposição de fatores entre 1998 a 2008. Cadernos de Saúde Pública, v. 34, p. e00213816, 2018.

BHARDWAJ, Arpit; TIWARI, Aruna. Breast cancer diagnosis using genetically optimized neural network model. Expert Systems with Applications, v. 42, n. 10, p. 4611-4620, 2015.

BRITO-SILVA, Keila; BEZERRA, Adriana Falangola Benjamin; TANAKA, Oswaldo Yoshimi. Direito à saúde e integralidade: uma discussão sobre os desafios e caminhos para sua efetivação.

CARUANA, Rich et al. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. 2015. p. 1721-1730.

CELEBI, M. Emre; AYDIN, Kemal (Ed.). Unsupervised learning algorithms. Berlin: Springer International Publishing, 2016.

CISMONDI, Federico et al. Missing data in medical databases: Impute, delete or classify?. Artificial intelligence in medicine, v. 58, n. 1, p. 63-72, 2013.

CHRISTINA, S. Shalin; SANTIAGO, Nirmala. Decision Support System for a Chronic Disease-Diabetes. 2018.

CROSS, Sarah H. et al. Rural-urban differences in cardiovascular mortality in the US, 1999-2017. Jama, v. 323, n. 18, p. 1852-1854, 2020.

DA SILVA, Ivan Nunes et al. Artificial neural network architectures and training processes. In: Artificial neural networks. Springer, Cham, 2017. p. 21-28.

DEO, Rahul C. Machine learning in medicine. Circulation, v. 132, n. 20, p. 1920-1930, 2015.

DESMET, Bart; HOSTE, Véronique. Online suicide prevention through optimised text classification. Information Sciences, v. 439, p. 61-78, 2018.

DOSHI-VELEZ, Finale; KIM, Been. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608, 2017.

ESTEVA, Andre et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature, v. 542, n. 7639, p. 115, 2017.

GAMBHIR, Shalini et al. Early Diagnostics Model for Dengue Disease Using Decision Tree-Based Approaches. Pre-Screening Systems for Early Disease Prediction, Detection, and Prevention, p. 69, 2018.

GIANFRANCESCO, Milena A. et al. Potential biases in machine learning algorithms using electronic health record data. JAMA internal medicine, v. 178, n. 11, p. 1544-1547, 2018.

HE, Jianxing et al. The practical implementation of artificial intelligence technologies in medicine. Nature medicine, v. 25, n. 1, p. 30-36, 2019.

HEARST, Marti A.. et al. Support vector machines. IEEE Intelligent Systems and their applications, v. 13, n. 4, p. 18-28, 1998.

JI, Shaoxiong et al. Supervised Learning for Suicidal Ideation Detection in Online User Content. Complexity, v. 2018, 2018.

JING, Chen; HOU, Jian. SVM and PCA based fault classification approaches for complicated industrial process. Neurocomputing, v. 167, p. 636-642, 2015.

JOSHI, Apurva; DANGRA, J.; RAWAT, M. A Decision Tree based classification technique for accurate heart disease classification and prediction. Int J Technol Res Manag, v. 3, p. 1-4, 2016.

KAELBLING, Leslie Pack; LITTMAN, Michael L.; MOORE, Andrew W. Reinforcement learning: A survey. Journal of artificial intelligence research, v. 4, p. 237-285, 1996.

KAUR, Gaganjot; CHHABRA, Amit. Improved J48 classification algorithm for the prediction of diabetes. International Journal of Computer Applications, v. 98, n. 22, 2014.

KAUR, Harsurinder; PANNU, Husanbir Singh; MALHI, Avleen Kaur. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. ACM Computing Surveys (CSUR), v. 52, n. 4, p. 1-36, 2019.

KIRBY, James B.; YABROFF, K. Robin. Rural–Urban Differences in Access to Primary Care: Beyond the Usual Source of Care Provider. American journal of preventive medicine, v. 58, n. 1, p. 89-96, 2020.

LI, Sheng; TANG, Bo; HE, Haibo. An imbalanced learning based MDR-TB early warning system. Journal of medical systems, v. 40, n. 7, p. 164, 2016.

LI, Jinyan et al. Adaptive swarm balancing algorithms for rare-event prediction in imbalanced healthcare data. PloS one, v. 12, n. 7, p. e0180830, 2017.

OH, Jihoon et al. Classification of suicide attempts through a machine learning algorithm based on multiple systemic psychiatric scales. Frontiers in psychiatry, v. 8, p. 192, 2017.

PAL, Shankho Subhra. Grey Wolf Optimization Trained Feed Foreword Neural Network for Breast Cancer Classification. International Journal of Applied Industrial Engineering (IJAIE), v. 5, n. 2, p. 21-29, 2018.

PAN, Ian et al. Machine learning for social services: a study of prenatal case management in Illinois. American journal of public health, v. 107, n. 6, p. 938-944, 2017.

PONTE, Francisco Diego Rabelo da. Preditores de heterogeneidade cognitiva no transtorno bipolar: uma abordagem machine-learning. 2018.

RATTAN, Sheenam et al. An optimized lung cancer classification system for computed tomography images. In: Image Information Processing (ICIIP), 2017 Fourth International Conference on. IEEE, 2017. p. 1-6.

ROFFMAN, David et al. Predicting non-melanoma skin cancer via a multi-parameterized artificial neural network. Scientific reports, v. 8, n. 1, p. 1701, 2018.

RIBEIRO, Marco Tulio; SINGH, Sameer; GUESTRIN, Carlos. " Why should i trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016. p. 1135-1144.

SÁEZ, José A.; KRAWCZYK, Bartosz; WOŹNIAK, Michał. Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. Pattern Recognition, v. 57, p. 164-178, 2016.

SANTOS, Hellen Geremias dos. Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina. 2018. Tese de Doutorado. Universidade de São Paulo.

SANTOS, Hellen Geremias dos et al. Machine learning para análises preditivas em saúde: exemplo de aplicação para predizer óbito em idosos de São Paulo, Brasil. Cadernos de Saúde Pública, v. 35, 2019.

SANTOS, Laércio Ives et al. Swarm intelligence and fuzzy sets for bed exit detection of elderly. Journal of Intelligent & Fuzzy Systems, v. 39, n. 1, p. 1061-1072, 2020.

SELVARAJU, Ramprasaath R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. 2017. p. 618-626.

SEMOLINI, R. Support vector machines, inferência transdutiva e o problema de classificação. 2002. 2002. Tese de Doutorado. Dissertação (Mestrado em Engenharia Elétrica)–Programa de Pósgraduação em Engenharia Elétrica, Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas, Campinas.

SOUZA, Maria de Fátima Marinho de et al. Transição da saúde e da doença no Brasil e nas Unidades Federadas durante os 30 anos do Sistema Único de Saúde. Ciência & Saúde Coletiva, v. 23, p. 1737-1750, 2018.

SILVA, Cleyton Cézar Souto et al. Artificial neural network and the decision support in nutritional food safety model. Journal of Nursing UFPE on line, v. 9, n. 3, p. 7078-7085, 2015.

SILVA, Fernando Henrique da et al. Estudo e desenvolvimento de métodos para predição de doadores de sangue. 2018.

SISODIA, Deepti; SISODIA, Dilip Singh. Prediction of diabetes using classification algorithms. Procedia computer science, v. 132, p. 1578-1585, 2018.

SUN, Yanmin et al. Cost-sensitive boosting for classification of imbalanced data. Pattern recognition, v. 40, n. 12, p. 3358-3378, 2007.

TORRES, Roberto L. Shinmoto et al. Sensor enabled wearable RFID technology for mitigating the risk of falls near beds. In: 2013 IEEE international conference on RFID (RFID). IEEE, 2013. p. 191-198.

VENKATASUBRAMANIAN, Venkat. Prognostic and diagnostic monitoring of complex systems for product lifecycle management: Challenges and opportunities. Computers & chemical engineering, v. 29, n. 6, p. 1253-1263, 2005.

VERANI, José Fernando de Souza; LAENDER, Fernando. A erradicação da poliomielite em quatro tempos. Cadernos de Saúde Pública, v. 36, p. e00145720, 2020.

VIDYA, M.; KARKI, Maya V. Skin cancer detection using machine learning techniques. In: 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT). IEEE, 2020. p. 1-5.

WARING, Jonathan; LINDVALL, Charlotta; UMETON, Renato. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. Artificial Intelligence in Medicine, v. 104, p. 101822, 2020.

WIENS, Jenna; SHENOY, Erica S. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. Clinical Infectious Diseases, v. 66, n. 1, p. 149-153, 2018.

YARDIMCI, Ahmet. Soft computing in medicine. Applied Soft Computing, v. 9, n. 3, p. 1029-1043, 2009.

YASSIN, Nisreen IR et al. Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. Computer methods and programs in biomedicine, v. 156, p. 25-45, 2018.

YU, Lingming et al. Prediction of pathologic stage in non-small cell lung cancer using machine learning algorithm based on CT image feature analysis. BMC cancer, v. 19, n. 1, p. 1-12, 2019.

ZOABI, Yazeed; DERI-ROZOV, Shira; SHOMRON, Noam. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. npj digital medicine, v. 4, n. 1, p. 1-5, 2021.

APÊNDICES

Apêndice A – Errata do Artigo 1

Página do Artigo 1	Descrição	Onde lia-se:	Leia-se:
5	Parágrafo 1 da seção 2.1.2	If p has both high hit rate and high concentration level, it will not move in in current iteration.	If p has both high hit rate and high concentration level, it will not move in current iteration.
6	Parágrafo 1 da seção 2.1.3	Figure 4(a) show	Figure 4(a) shows
6	Parágrafo 2 da seção 2.1.3	To get best results, the NcPSC uses 6 particles for classifing classify the classes	To get best results, the NcPSC uses 6 particles for classifing the classes
6	Parágrafo 4 da seção 2.2	when the arithmetic mean of sliding window W3 is greater than arithmetic mean of the other two classes.	when the arithmetic mean of sliding window W3 is greater than arithmetic mean of the other two classes. In Eq. (7), W_{ij} represents the jth value in the ith window.
8	Tabela 3	Number of subjects	Number of Participants
8	Tabela 3	Number of Number of observations	Number of observations
8	Parágrafo 2 da seção 3.2	Finally, was used the f-score (Eq. 10) to calculate the amount of FN and FP.	Finally, the f-score (Eq. 10) was used to calculate the harmonic mean between precision and recall. the f-score is efficient to assess the quality of the approach in unbalanced datasets, such as the dataset used in this work.
9	Parágrafo 1	Then, was used the other participants, separately, to test the model.	Then the other participants was used, separately, to test the model
9	Parágrafo 2	In the same way, in room 2, present in Table 5, it can see that both are very close.	In the same way, the results for room 2 are presented in Table 5, and it can be seen that both are very close.
10	Parágrafo 3	Finally, the main advantages of the proposed approach is discussed:	Finally, the main advantages of the proposed approach are discussed:
11	Referência número [3]	Caminhas,W.M., Takahashi, R.H.: Dynamic system failure detection and diagnosis employing sliding mode observers and fuzzy neural networks. In: Proceedings Joint 9th IFSA World Congress and 20th NAFIPS International Conference (Cat. No.01TH8569), vol. 1, pp. 304–309. IEEE (2001)	SANTOS, Laércio I. et al. A new scheme for fault detection and classification applied to dc motor. TEMA (São Carlos), v. 19, p. 327-345, 2018.