

Identificação e Análise de Dias Depressivos a Partir da Classificação em Séries Temporais de Atividade Motora

Guilherme A. Rocha de Figueiredo¹, Julio C. S. Reis¹

¹Departamento de Informática – Universidade Federal de Viçosa (UFV) – Brasil
guilherme.figueiredo@ufv.br, jreis@ufv.br

Abstract. *This study examines the use of various time series classifiers to identify depressed days by analyzing daily motor activity in individuals with and without a clinical diagnosis of the condition. The results show a significant improvement in time series algorithms compared to baseline tabular models. However, accurately detecting depressed days remains challenging. Qualitative error analysis indicates that failures are consistent across different algorithmic families and are associated with specific circadian profiles influenced by external factors. Nonetheless, the technique shows promise as a foundation for longitudinal monitoring, supporting the development of effective public health policies in systems such as the Brazilian Unified Health System (SUS), and complementing the work of specialized professionals.*

Resumo. *Este estudo investiga o uso de diferentes classificadores de séries temporais para a identificação de dias depressivos por meio da análise da atividade motora diária de indivíduos com e sem diagnóstico clínico da doença. Os resultados mostram uma melhora significativa nos algoritmos de séries temporais em comparação com os modelos tabulares de referência. No entanto, detectar com precisão os dias de depressão continua sendo um desafio. A análise qualitativa de erros indica que as falhas são consistentes entre diferentes famílias de algoritmos e estão associadas a perfis circadianos específicos influenciados por fatores externos. Ainda assim, a técnica se mostra promissora como base para o monitoramento longitudinal, apoiando o desenvolvimento de políticas de saúde pública eficazes em sistemas como o Sistema Único de Saúde (SUS) brasileiro e complementando o trabalho de profissionais especializados.*

1. Introdução

A depressão é um grave problema de saúde pública global. Dados da Organização Mundial da Saúde (OMS) indicam que mais de 300 milhões de pessoas sofrem com o transtorno em todo o mundo, sendo o Brasil o país com a maior prevalência de depressão na América Latina [World Health Organization 2017]. Além do imenso impacto na qualidade de vida, essas patologias resultam em uma perda massiva de produtividade, gerando um prejuízo anual estimado em US\$ 1 trilhão para a economia global¹. Esse cenário foi significativamente agravado pela pandemia de COVID-19; estima-se que, apenas no primeiro ano da crise sanitária, houve um aumento de 25% na prevalência global de ansiedade e depressão [World Health Organization 2022].

Apesar dessa gravidade, o diagnóstico e o monitoramento da depressão enfrentam obstáculos significativos. Tradicionalmente, a avaliação clínica depende de entrevistas

¹<https://news.un.org/pt/story/2016/04/1547361>

esporádicas e relatos subjetivos, suscetíveis a vieses de memória e percepção [American Psychiatric Association et al. 2013]. Para mitigar essa subjetividade, a comunidade científica tem investido na chamada fenotipagem digital: o uso de dispositivos comuns do dia a dia, como *smartphones* e relógios inteligentes (*wearables*), para coletar sinais digitais passivos de forma contínua e quantificar sintomas físicos intrínsecos aos transtornos de humor. A análise automatizada desses dados por algoritmos de aprendizado de máquina (ML do inglês, *Machine Learning*) possibilita a detecção de padrões comportamentais associados a episódios depressivos de forma objetiva [Benoit et al. 2020].

Neste contexto, recentemente, os autores em [Song et al. 2025] revisaram abordagens de ML para detectar depressão a partir de sinais digitais, apontando limitações críticas como a escassez de dados rotulados e públicos, a baixa explicabilidade dos modelos, e o viés de classe. Este trabalho busca mitigar tais lacunas ao avaliar diferentes abordagens de classificadores de séries temporais para o problema, preservando a comparabilidade entre métodos enquanto investiga detalhadamente as instâncias de erro para elevar a interpretabilidade dos resultados. Especificamente, este estudo busca responder às seguintes questões de pesquisa (QPs): *QP1) Qual o desempenho das diferentes famílias de classificadores de séries temporais na detecção de dias depressivos em comparação a abordagens tabulares?*; *QP2) Quais as limitações e os perfis de erro associados à classificação de transtornos mentais em janelas diárias estritas?*

Para responder a essas questões, explorou-se o conjunto de dados *Depression* [Garcia-Ceja et al. 2018a], que contém registros contínuos de actigrafia² de indivíduos saudáveis e diagnosticados com depressão. Investigou-se ainda o potencial de seis abordagens representativas de famílias distintas de séries temporais, comparando-as a modelos tradicionais. Em resumo, os resultados indicam que abordagens baseadas em séries temporais superam a tabular em métricas globais, embora persista o desafio em capturar dias depressivos atípicos devido ao ruído biológico dos rótulos. Ademais, a análise de erros revela que falhas algorítmicas tendem a ser consensuais e ligadas a picos de atividade motora atípicos, sugerindo que o monitoramento longitudinal é essencial para políticas públicas de saúde digital mais assertivas.

O restante do trabalho está organizado da seguinte forma. A Seção 2 discute trabalhos correlatos. Depois, a Seção 3 descreve a metodologia aplicada. A Seção 4 apresenta os desempenhos globais dos modelos, enquanto a Seção 5 aprofunda a análise de falhas através de experimentos específicos. Por fim, a Seção 6 traz as conclusões e diretrizes para trabalhos futuros.

2. Trabalhos Relacionados

A utilização de actígrafos para a avaliação de transtornos de humor tem sido amplamente documentada na literatura médica e computacional. O estudo conduzido por Tazawa et al. [Tazawa et al. 2019] apresenta uma revisão sistemática de trabalhos neste contexto e atesta a existência de diferenças objetivas e significativas entre o padrão de comportamento motor de pacientes psiquiátricos e indivíduos saudáveis. No entanto, os autores reconhecem limitações inerentes à área, como bases de dados reduzidas e a influência de variáveis não controladas, incluindo medicação e hospitalização.

²A actigrafia é um método não invasivo para monitorar o ciclo de atividade e repouso humano, utilizando sensores de pulso para registrar o movimento motor de forma contínua [Benedito-Silva et al. 2021].

No contexto computacional, a aplicação de modelos preditivos sobre esses sinais digitais apresenta grande potencial. Um levantamento abrangente realizado por Garcia-Ceja *et al.* [Garcia-Ceja et al. 2018b] mapeia a crescente adoção de técnicas de aprendizado de máquina processando dados multimodais advindos de sensores *wearables* para o monitoramento contínuo em saúde mental. A principal vantagem deste tipo de abordagem reside na coleta não invasiva e ecológica, refletindo o paciente em seu ambiente natural.

Foi nesse cenário de interseção tecnológica e clínica que Garcia-Ceja *et al.* [Garcia-Ceja et al. 2018a] introduziram o conjunto de dados *Depresjon*, utilizado no presente estudo. No trabalho original, os autores forneceram uma linha de base (*baseline*) de classificação utilizando algoritmos tabulares tradicionais alimentados por estatísticas diárias simples. Desde a sua publicação, este conjunto de dados tem sido explorado sob diferentes óticas computacionais. Em [Aminifar et al. 2021], por exemplo, os autores empregaram técnicas de aprendizado distribuído com preservação de privacidade e aumento artificial de dados, alcançando uma acurácia de 76,8%. Sob outra perspectiva, Espino-Salinas *et al.* [Espino-Salinas et al. 2022] obtiveram um desempenho semelhante (76,7%) ao modelar o problema por meio de Redes Neurais Convolucionais 2D (2D-CNN), convertendo as séries temporais em matrizes bidimensionais (dia vs. semana) em que os registros de atividade por minuto são processados como pixels de uma imagem.

O presente trabalho se difere e expande esses estudos anteriores ao focar no paradigma das séries temporais em sua forma pura. Ao preservar a dimensionalidade natural dos dados (minuto a minuto) e testar diferentes famílias de classificação temporal, busca-se não apenas superar as linhas de base, mas, sobretudo, investigar qualitativamente a origem das falhas algorítmicas frente à volatilidade comportamental da depressão.

3. Metodologia

Nesta seção, apresenta-se a metodologia experimental desenvolvida para este trabalho. Com vistas à reprodutibilidade e à comparação justa de resultados, a abordagem adotada seguiu um padrão metodológico semelhante ao do artigo original que introduziu o conjunto de dados [Garcia-Ceja et al. 2018a].

3.1. Conjunto de Dados

Neste trabalho, explorou-se o conjunto de dados público *Depresjon* [Garcia-Ceja et al. 2018a]. Esta base é composta por registros contínuos da atividade motora de 55 indivíduos, coletados ao longo de pelo menos duas semanas utilizando um actígrafo de pulso no braço direito. Os participantes dividem-se em quatro grupos diagnósticos distintos: (i) depressão unipolar, (ii) transtorno bipolar tipo I, (iii) transtorno bipolar tipo II e (iv) um grupo controle formado por indivíduos saudáveis.

No formato original, cada série temporal possui comprimentos variados, com a atividade motora registrada minuto a minuto. Para fins de padronização temporal, visando facilitar o processamento algorítmico e aumentar o volume de amostras para o treinamento, as séries contínuas de cada paciente foram segmentadas em janelas diárias estritas de 1.440 minutos. Os dias com registros incompletos foram descartados, resultando em um volume total de 1.029 séries temporais (dias).

A Figura 1 ilustra a distribuição desses 1.029 dias entre as subclasses. Observa-se que as 14 amostras relativas ao transtorno bipolar tipo I são escassas e provenientes de um

único paciente. Para evitar que os algoritmos de aprendizado de máquina memorizassem os padrões motores específicos desse indivíduo (*patient-level bias*), essa subclasse foi removida do experimento realizado e apresentado neste estudo.

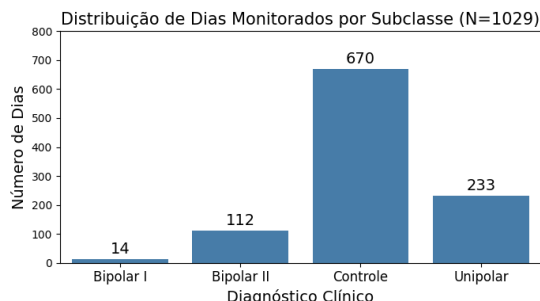


Figura 1. Distribuição das séries temporais (dias) entre as diferentes subclasses clínicas.

do grupo patológico estavam diagnosticados e monitorados durante o período, não é possível afirmar de forma categórica que todos os dias avaliados manifestaram sintomas depressivos agudos, devido à natural flutuação de humor e energia da doença. Diante disso, este trabalho também se propõe a investigar o impacto desse ruído biológico na eficácia da detecção.

3.2. Modelos de Classificação

Este trabalho explora modelos de classificação de séries temporais baseados em diferentes arquiteturas, com o intuito de investigar a representação do problema em múltiplos paradigmas computacionais. Para isso, foram selecionados seis algoritmos específicos para séries temporais, cada um atuando como representante canônico de uma família (ou taxonomia) estabelecida pelo referencial estudo comparativo (*bake-off*) apresentado em [Bag-nall et al. 2017], e recentemente expandido em [Middlehurst et al. 2024].

A Tabela 1 lista os algoritmos selecionados, suas respectivas famílias taxonômicas e uma breve descrição do funcionamento de cada um deles. Além dos modelos voltados a séries temporais, adotou-se o classificador tabular *Linear Support Vector Classification* (Linear SVC) como linha de base (*baseline*). Esse algoritmo apresentou o melhor desempenho na maioria das métricas no estudo original do conjunto *Depresjon* [Garcia-Ceja et al. 2018a], e foi replicado neste trabalho utilizando o mesmo vetor de características estatísticas: média, desvio padrão e proporção de zeros (inatividade) da atividade motora.

3.3. Configuração Experimental

Para realização dos experimentos, aplicou-se a validação cruzada estratificada com 10 partições (*Stratified 10-Fold Cross-Validation*), preservando a proporção real de classes em cada iteração e garantindo a geração de uma predição para cada instância da base. Adicionalmente, para mitigar vieses de escala, aplicou-se o *MinMaxScaler*³ estritamente ao conjunto de treinamento de cada partição, replicando-se os mesmos parâmetros de normalização para o respectivo conjunto de teste.

³<https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html>

O restante dos dados (1.015 instâncias válidas) foi remapeado para um problema de classificação binária: 670 dias foram categorizados como não depressivos (advindos do grupo controle), e 345 dias foram categorizados como potencialmente depressivos (agrupando os 233 dias de depressão unipolar e os 112 dias de bipolar tipo II). Uma limitação intrínseca desta abordagem de segmentação temporal é o viés de rotulação (*label noise*). Embora haja a garantia clínica de que os pacientes

Tabela 1. Taxonomia e descrição dos classificadores de séries temporais.

Classificador	Família	Descrição
<i>Time Series Forest</i>	Baseado em Intervalos	Extraí características a partir de sub-intervalos aleatórios.
Arsenal	Baseado em Convolução	Aplica filtros convolucionais (núcleos/kernels) aleatórios para capturar padrões.
CBOSS	Baseado em Dicionário	Converte a série temporal em "palavras" e utiliza o histograma de frequência desses padrões para classificação.
Catch22	Baseado em Características	Extraí 22 características globais do sinal temporal e as encaminha para um classificador padrão.
KNN com DTW	Baseado em Distância	Classifica utilizando o vizinho mais próximo, com a distancia sendo a similaridade (<i>Dynamic Time Warping</i>) entre as séries temporais.
<i>Shapelet Transform</i>	Baseado em Subsequência	Procura por subsequências curtas (<i>shapelets</i>) com alta capacidade discriminatória entre as classes.

Para avaliar o desempenho dos classificadores, foram exploradas as mesmas métricas reportadas no estudo seminal: Acurácia, Precisão, Sensibilidade (*Recall*), Especificidade, *F1-Score* e o Coeficiente de Correlação de Matthews (MCC). O cálculo dessas métricas foi realizado de forma consolidada (*Out-of-Fold*). Nesta abordagem, as predições de todas as instâncias são coletadas antes do cálculo global dos indicadores de desempenho, fornecendo uma estimativa mais robusta da performance dos algoritmos.

Os experimentos foram implementados em Python utilizando as bibliotecas *scikit-learn* e *sktime*, e executados em um computador com um processador AMD Ryzen 5 5600H e 16GB de memória RAM. Visando avaliar a capacidade de generalização nativa (*out-of-the-box*) das arquiteturas e garantir a reprodutibilidade, manteve-se os hiperparâmetros padrão das bibliotecas, alterando-se estritamente o parâmetro de balanceamento de classes (*class_weight*), quando disponível, para compensar a desigualdade no número de amostras. O código-fonte está publicamente disponível em: <https://github.com/GuiFigueiredo0/minds-workshop-depresjon>.

4. Resultados de Classificação

Para responder à QP1 referente ao desempenho computacional, a Tabela 2 apresenta os resultados gerais dos diferentes modelos de classificação. Nota-se que praticamente todos os algoritmos baseados em séries temporais melhoraram as médias ponderadas para todas as métricas, com exceção da especificidade para os algoritmos Catch22 e CBOSS. Ao analisar os resultados das métricas para cada uma das subclasses, fica evidente que esse decréscimo ocorreu porque tais algoritmos sofreram um sobreajuste (*overfitting*) à classe predominante (controle). É importante ressaltar o desempenho do *Time Series Forest* (TSF), que obteve os melhores valores globais (80,3%), superou não apenas a linha de base tabular, mas também os resultados em torno de 76% reportados em estudos anteriores que exploraram o mesmo conjunto de dados [Aminifar et al. 2021, Espino-Salinas et al. 2022]. Esse resultado revela como a representação contínua em séries temporais permite uma análise mais completa do padrão motor pelo algoritmo, elevando seu poder de classificação.

Além disso, foi possível identificar uma limitação persistente nos algoritmos, evidenciada principalmente pelas métricas de sensibilidade e especificidade: os modelos conseguiram melhorar a identificação dos indivíduos saudáveis (dias não depressivos), com o TSF melhorando a sensibilidade para classe de controle em 20%, mas a detecção de dias depressivos ainda se apresenta como uma tarefa altamente desafiadora. A habilidade de classificar corretamente os dias patológicos manteve-se estagnada mesmo na

Tabela 2. Resultados do desempenho dos classificadores.

Classificador	Subclasse	Precisão	Sensibilidade	Acurácia	MCC	F1-Score	Especificidade
SVM Tabular	Depressivo	0.5124	0.6580	0.6709	0.3205	0.5761	0.6776
	Não Depressivo	0.7937	0.6776	0.6709	0.3205	0.7311	0.6580
	Média Ponderada	0.6981	0.6709	0.6709	0.3205	0.6784	0.6646
TSF	Depressivo	0.7346	0.6580	0.8030	0.5513	0.6942	0.8776
	Não Depressivo	0.8329	0.8776	0.8030	0.5513	0.8547	0.6580
	Média Ponderada	0.7995	0.8030	0.8030	0.5513	0.8001	0.7326
Arsenal	Depressivo	0.6898	0.6058	0.7734	0.4818	0.6451	0.8597
	Não Depressivo	0.8090	0.8597	0.7734	0.4818	0.8336	0.6058
	Média Ponderada	0.7685	0.7734	0.7734	0.4818	0.7695	0.6921
Catch22	Depressivo	0.6628	0.4957	0.7429	0.3980	0.5672	0.8701
	Não Depressivo	0.7701	0.8701	0.7429	0.3980	0.8171	0.4957
	Média Ponderada	0.7337	0.7429	0.7429	0.3980	0.7321	0.6229
CBOSS	Depressivo	0.7747	0.4087	0.7586	0.4291	0.5351	0.9388
	Não Depressivo	0.7551	0.9388	0.7586	0.4291	0.8370	0.4087
	Média Ponderada	0.7618	0.7586	0.7586	0.4291	0.7344	0.5889
KNN DTW	Depressivo	0.5318	0.6551	0.6867	0.3438	0.5870	0.7030
	Não Depressivo	0.7983	0.7030	0.6867	0.3438	0.7476	0.6551
	Média Ponderada	0.7077	0.6867	0.6867	0.3438	0.6930	0.6714
Shapelet	Depressivo	0.7416	0.5739	0.7872	0.5066	0.6471	0.8970
	Não Depressivo	0.8035	0.8970	0.7872	0.5066	0.8477	0.5739
	Média Ponderada	0.7824	0.7872	0.7872	0.5066	0.7795	0.6837

transição do melhor algoritmo tabular (linha de base) para o melhor modelo de séries temporais (TSF). Em direção à QP2, foram executados experimentos qualitativos, descritos a seguir, a fim de elucidar as limitações e perfis de erro associados às janelas diárias.

4.1. Consenso de Falhas

A primeira análise diz respeito ao consenso de falhas dos algoritmos, cujo objetivo é investigar se os modelos estão classificando incorretamente as mesmas instâncias (dias) ou se cada um apresenta um viés de erro distinto. Essa análise é particularmente relevante neste contexto, visto que os classificadores executados pertencem a famílias matemáticas e estruturais intrinsecamente diferentes. A Figura 2 ilustra a quantidade de classificadores que falharam na predição de um mesmo dia, segmentada por classe (depressão e controle).

Os resultados apresentados nesta figura evidenciam que distribuição de falhas para a classe não depressiva exhibe o declínio esperado: há uma probabilidade maior de uma amostra ser classificada erroneamente por um ou dois algoritmos, mas as chances de falha simultânea por todos os modelos caem drasticamente. Contudo, a classe de depressão manteve uma taxa de falha quase estática e linear à medida que o número de algoritmos aumentou. Consequentemente, esse cenário inviabiliza melhorias significativas através da implementação de técnicas de comitê (*Ensemble*) [Zhou 2025], nas quais a classificação final de uma amostra é definida pela maioria dos votos dos classificadores base. Como a premissa de um *Ensemble* de sucesso exige diversidade nos erros dos modelos componentes, a falha unânime nas mesmas amostras depressivas torna a votação redundante.

5. Análise Qualitativa e Comportamental dos Erros

Para aprofundar a investigação dos resultados globais, as análises a seguir utilizam exclusivamente as predições geradas pelo modelo *Time Series Forest* (TSF). Esta escolha

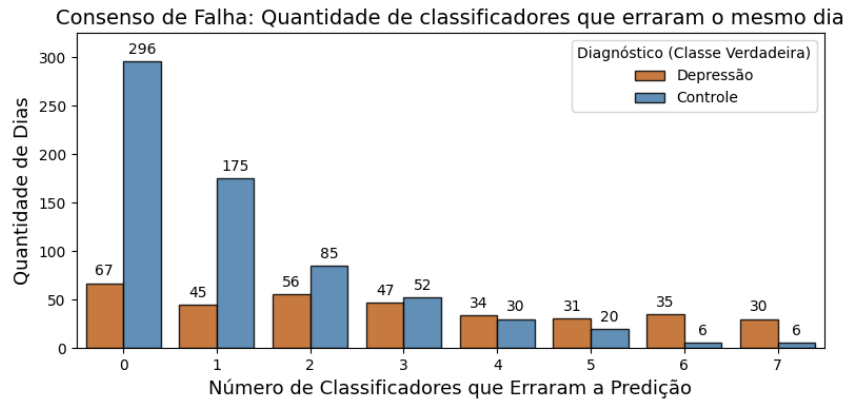


Figura 2. Quantidade de classificadores que erraram o mesmo dia, separado por classe. O valor zero no eixo X representa amostras corretamente classificadas por todos os classificadores, enquanto 7 representa amostras classificadas erroneamente por todos os classificadores.

justifica-se pelo fato de o TSF ter apresentado o melhor desempenho geral e, especificamente, a melhor sensibilidade para a classe depressiva entre as abordagens investigadas.

5.1. Confiância do Modelo

Antes de proceder à análise qualitativa, é imperativo investigar a confiabilidade das predições do TSF. Para tanto, realizou-se o reprocessamento do modelo com a fixação da semente aleatória (*seed*) original, configurando-o para extrair as probabilidades de classificação em vez de apenas os rótulos binários. Esses dados fundamentaram a construção da curva ROC e do gráfico de distribuição de confiança.

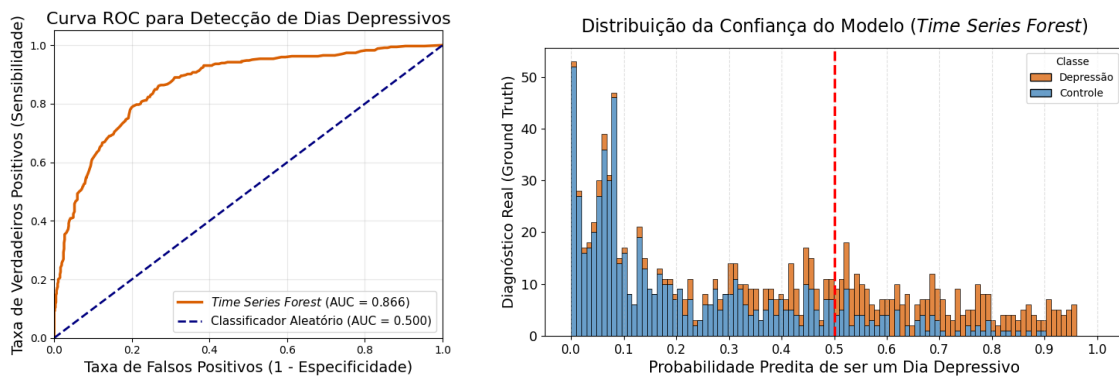


Figura 3. Métricas de decisão (curva ROC) – à esquerda; e análise de probabilidade - à direita. Ambos para o modelo TSF.

A Figura 3.1 apresenta a curva ROC com uma Área Sob a Curva (AUC) de 0,866, valor considerado robusto sob a ótica computacional. A Figura 3.2 corrobora este achado, evidenciando que o algoritmo demonstra alta convicção na classificação da maioria das instâncias, com os erros concentrando-se majoritariamente na zona de incerteza próxima ao limiar de 0,5. Este comportamento reflete a ambiguidade diagnóstica inerente ao mundo real, onde a transição entre padrões motores saudáveis e patológicos não é discreta, mas sim uma zona de sobreposição clínica.

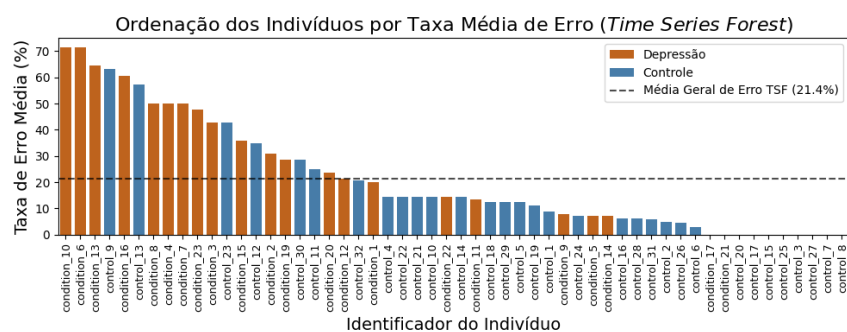


Figura 4. Ordenação de todos os indivíduos por taxa de erro média (TSF).

5.2. Taxa Média de Erro por Indivíduo

Esta análise agrupa as séries temporais diárias de volta aos seus respectivos indivíduos de origem, considerando os 54 pacientes presentes no conjunto de dados final. O objetivo é identificar se as falhas do modelo são distribuídas aleatoriamente por toda a base ou se estão concentradas em pacientes específicos. A Figura 4 ordena todos os indivíduos avaliados de acordo com a sua respectiva taxa de erro média.

Observa-se que a distribuição de erros é fortemente heterogênea. Enquanto o grupo controle foi classificado com precisão (a maioria com erros próximos a zero), as falhas mais acentuadas residem em indivíduos da classe depressiva. A concentração de taxas de erro próximas a 50% para certos indivíduos reforça a ambiguidade diagnóstica do dado. Levanta-se a hipótese de que essa dificuldade computacional possa estar refletindo o fenômeno clínico da “máscara social”, onde o paciente mimetiza um padrão funcional saudável, ocultando sinais objetivos da doença [Bril-Barniv et al. 2017].

5.3. Perfil de Atividade Motora Circadiano

Nesta última etapa, investiga-se a variação da atividade motora ao longo do dia com base no comportamento de acertos e erros do modelo. Primeiramente, as amostras diárias foram separadas em quatro subgrupos, definidos pelo tipo de predição gerada pelo TSF em relação ao diagnóstico real: Verdadeiros Positivos (TP), Verdadeiros Negativos (TN), Falsos Positivos (FP) e Falsos Negativos (FN). Em seguida, a atividade motora dessas instâncias foi agrupada e calculou-se a média hora a hora (de 0h às 23h). A Figura 5 apresenta o resultado dessa agregação, exibindo a média e o intervalo de confiança (95%) do ritmo circadiano segmentado por tipo de predição.

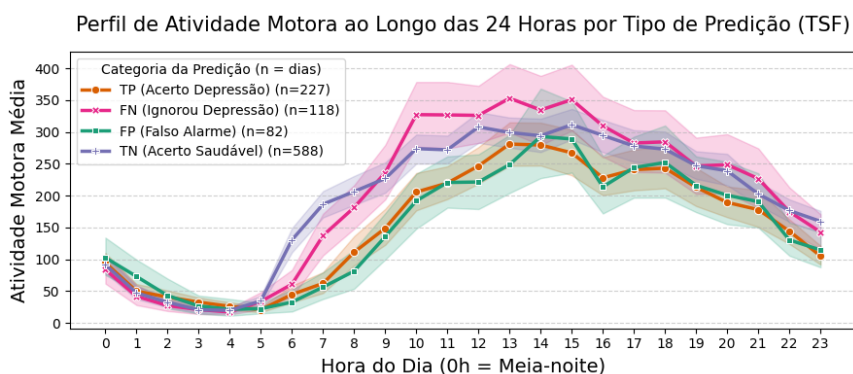


Figura 5. Perfil de atividade motora diário (agregado) segmentado por tipo de predição do TSF.

Tomando o comportamento do grupo controle (TN) como base normativa, nota-se que os dias depressivos corretamente identificados (TP) apresentam níveis de atividade significativamente reduzidos. Esse padrão reflete o retardo psicomotor e a letargia, que são sintomas cardinais e altamente característicos da depressão clínica [Buyukdura et al. 2011]. Em contrapartida, os Falsos Negativos (FN) exibem atividade motora superior à norma e maior variabilidade. Especula-se que tal fenômeno possa estar associado a fatores externos (como demandas laborais) ou possíveis comorbidades não documentadas nos metadados, como ansiedade, que frequentemente introduzem estados de agitação, mascarando a redução de mobilidade esperada no quadro depressivo clássico.

6. Conclusão e Trabalhos Futuros

Este trabalho avaliou o desempenho de diferentes abordagens de séries temporais na detecção de dias depressivos, visando responder às questões de pesquisa (QPs) propostas inicialmente. Em resposta à QP1, os resultados demonstraram que os algoritmos baseados em séries temporais superaram a linha de base tabular em métricas globais, sugerindo que a representação contínua da actigrafia pode oferecer uma extração de padrões mais rica. Entretanto, observou-se que o ganho em acurácia global não se traduziu integralmente em uma melhora na sensibilidade para a classe patológica, indicando que a arquitetura do modelo é apenas um dos componentes do desafio.

No que tange à QP2, referente às limitações e perfis de erro em janelas diárias estritas, a análise qualitativa indicou que a principal barreira parece não residir apenas em limitações computacionais, mas na complexidade biológica e no ruído de rótulo. O alto consenso de falhas entre algoritmos de diferentes famílias sugere que certas instâncias são inerentemente ambíguas quando analisadas isoladamente em um período de 24 horas. Como trabalhos futuros, sugere-se a exploração de abordagens híbridas que apliquem técnicas de redução de dimensionalidade sobre as séries temporais. Essa via pode atuar como um meio-termo promissor, buscando capturar a complexidade da dinâmica motora ao mesmo tempo em que filtra potenciais ruídos. Adicionalmente, técnicas de aumento artificial de dados podem ser aliadas a análises longitudinais para suprir a limitação volumétrica.

A análise aprofundada dos erros do modelo *Time Series Forest* (TSF) permitiu levantar hipóteses sobre as raízes dessas falhas na classe minoritária, complementando os achados da QP2. O algoritmo identificou com precisão o padrão clássico de letargia, mas falhou sistematicamente em dias marcados por agitação psicomotora, possivelmente induzida por ansiedade ou demandas laborais externas. Esse fenômeno apoia a tese de que a depressão é uma doença heterogênea cujas manifestações físicas podem mimetizar o padrão saudável em janelas temporais curtas, gerando a “borda incerta” discutida na hipótese deste estudo.

Por fim, ressalta-se que o propósito deste classificador é atuar de forma complementar, e não substitutiva, à avaliação psiquiátrica. Diante da popularização de *wearables*, ecossistemas de saúde pública como o Sistema Único de Saúde (SUS) poderiam integrar esses dados para monitoramento populacional. Embora o sinal motor de um único dia possa ser insuficiente para um diagnóstico definitivo, a persistência de alertas ao longo de janelas temporais estendidas serve como um marcador precoce valioso, auxiliando na formulação de políticas públicas mais assertivas e na identificação proativa de indivíduos que necessitam de suporte especializado.

Referências

- American Psychiatric Association, D., American Psychiatric Association, D., et al. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5*, volume 5. American Psychiatric Association.
- Aminifar, A., Rabbi, F., Pun, V. K. I., and Lamo, Y. (2021). Monitoring motor activity data for detecting patients' depression using data augmentation and privacy-preserving distributed learning. In *Annual Int'l Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2163–2169.
- Bagnall, A., Lines, J., Bostrom, A., Large, J., and Keogh, E. (2017). The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data mining and Knowledge Discovery*, 31(3):606–660.
- Benedito-Silva, A. A., Gonçalves, B., Vartanian, D., Santos, E. H. R., Miguel, M. A. L., Pedrazzoli, M., and Gonçalves, B. (2021). Especificações técnicas e recomendações sobre o uso da actigrafia. *Consenso Brasileiro de Actigrafia. São Paulo: Segmento Farma Editores*, pages 13–28.
- Benoit, J., Onyeaka, H., Keshavan, M., and Torous, J. (2020). Systematic review of digital phenotyping and machine learning in psychosis spectrum illnesses. *Harvard review of psychiatry*, 28(5):296–304.
- Bril-Barniv, S., Moran, G. S., Naaman, A., Roe, D., and Karnieli-Miller, O. (2017). A qualitative study examining experiences and dilemmas in concealment and disclosure of people living with serious mental illness. *Qualitative Health Research*, 27(4):573–583.
- Buyukdura, J. S., McClintock, S. M., and Croarkin, P. E. (2011). Psychomotor retardation in depression: biological underpinnings, measurement, and treatment. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 35(2):395–409.
- Espero-Salinas, C. H., Galván-Tejada, C. E., Luna-García, H., Gamboa-Rosales, H., Celaya-Padilla, J. M., Zanella-Calzada, L. A., and Tejada, J. I. G. (2022). Two-dimensional convolutional neural network for depression episodes detection in real time using motor activity time series of depresjon dataset. *Bioengineering*, 9(9):458.
- Garcia-Ceja, E., Riegler, M., Jakobsen, P., Tørresen, J., Nordgreen, T., Oedegaard, K. J., and Fasmer, O. B. (2018a). Depresjon: a motor activity database of depression episodes in unipolar and bipolar patients. In *Proc. of the ACM Multimedia Systems Conference*, pages 472–477.
- Garcia-Ceja, E., Riegler, M., Nordgreen, T., Jakobsen, P., Oedegaard, K. J., and Tørresen, J. (2018b). Mental health monitoring with multimodal sensing and machine learning: A survey. *Pervasive and Mobile Computing*, 51:1–26.
- Middlehurst, M., Schäfer, P., and Bagnall, A. (2024). Bake off redux: a review and experimental evaluation of recent time series classification algorithms: M. middlehurst et al. *Data Mining and Knowledge Discovery*, 38(4):1958–2031.
- Song, S., Huo, Y., Tang, S., Cheong, J., Gao, R., Valstar, M., and Gunes, H. (2025). Automatic depression assessment using machine learning: A comprehensive survey. *arXiv preprint arXiv:2506.18915*.
- Tazawa, Y., Wada, M., Mitsukura, Y., Takamiya, A., Kitazawa, M., Yoshimura, M., Mimura, M., and Kishimoto, T. (2019). Actigraphy for evaluation of mood disorders: A systematic review and meta-analysis. *Journal of Affective Disorders*, 253:257–269.
- World Health Organization (2017). Depression and other common mental disorders: global health estimates. Geneva. Licence: CC BY-NC-SA 3.0 IGO.
- World Health Organization (2022). World mental health report: Transforming mental health for all. Technical report, World Health Organization, Geneva.
- Zhou, Z.-H. (2025). *Ensemble methods: foundations and algorithms*. Chapman and Hall/CRC.