Enhancing dengue time-series forecasting at the neighborhood level based on the intensity of contagion

Rafael Bomfim[1,2], **J. L. B. de Araújo**[1], **Antonio S. Lima Neto**[4], **Vasco Furtado**[1,2,3]

[1]Laboratório de Ciência de Dados e Inteligência Artificial, Universidade de Fortaleza
Fortaleza, Ceará, 60811-905, Brasil

[2]Programa de Pós Graduação em Informática Aplicada, Universidade de Fortaleza
Fortaleza, Ceará, 60811-905, Brasil

[3]Empresa de Tecnologia da Informação do Ceará, Governo do Estado do Ceará
Fortaleza, Ceará, 60130-240, Brasil

[4]Célula de Vigilância Epidemiológica, Secretaria Municipal da Saúde
Fortaleza, Ceará, 60810-670, Brasil

{bomfim,jorgearaujo}@unifor.br, tanta26@yahoo.com, vasco@unifor.br

Abstract. *This article shows how models based on LSTM can be more accurate in the task of predicting Dengue cases by neighborhood as well as identifying neighborhoods that are more susceptible to an epidemic. It is suggested the use of models that incorporate heuristic information that capture the intensity of disease propagation normalized by the population. Models based on LSTM, with and without the proposed heuristic, are evaluated for their ability to predict cases for the entire city as well as for all neighborhoods. The improvement obtained when using such a strategy is highlighted.*

Resumo. *Este artigo mostra como modelos baseados em LSTM podem ter melhor acurácia na tarefa de prever casos de Dengue por bairro bem como identificar bairros mais suscetíveis de epidemia. Sugere-se o uso de modelos que incorporem informações heurísticas que capturam a intensidade de propagação da doença normalizada pela população. Os modelos baseados em LSTM, com e sem a heurística proposta, são avaliados quanto à capacidade de prever os casos para toda a cidade bem como para todos os bairros. Destaca-se a melhoria obtida ao se usar tal estratégia.*

1 Introdução

A ocorrência de doenças infecciosas em cidades varia no espaço e no tempo, pois depende das características socioambientais de cada região da cidade. Construir modelos preditivos nesse contexto requer mapear os impactos dessas diversas características que podem influenciar a transmissibilidade, tais como a população de um bairro ou o deslocamento de pessoas entre bairros [1, 2].

Para capturar a não-uniformidade do processo de contágio, modelos preditivos para doenças infecciosas devem considerar as especificidades das regiões da cidade, mas também as interações entre elas. A natureza multifatorial desse processo é desafiadora de representar, principalmente devido à falta de dados precisos e consistentemente atualizados. No caso da infecção por dengue, por exemplo, uma doença transmitida por um vetor

(no caso, o mosquito Aedes aegypti), a taxa de infestação de mosquitos em cada setor censitário ou mesmo bairro da cidade nem sempre está disponível [3, 4]. Além disso, os dados sobre a dinâmica de movimentação das pessoas nem sempre são acessíveis, embora o uso de informações de mobilidade do transporte público para esse fim tenha sido cada vez mais frequente [5, 6].

A literatura científica é abundante em modelos estatísticos e de aprendizado de máquina para previsão epidemiológica. Geralmente, esses modelos usam exclusivamente as séries temporais passadas de casos da doença. O modelo ARIMA é amplamente utilizado, como demonstrado em [7], onde previu a taxa de positividade do vírus Influenza em crianças em Wuhan. Em [8], os autores propuseram uma abordagem de rede neural (na sigla em inglês, NN), emulando o comportamento de um modelo de autorregressão para a previsão semanal de doenças semelhantes à influenza (ILI) nos EUA. O modelo de rede neural superou dois outros modelos de autorregressão, ARIMA e ARIMA-STL, em termos de precisão. Em [9], os autores introduziram uma rede neural baseada em LSTM capaz de capturar o impacto da mobilidade humana na evolução dos casos de dengue, produzindo melhores resultados em comparação com modelos compartimentais como o SIR. LSTMs também foram usadas para prever casos de COVID-19 [11] e Influenza [12].

Os trabalhos acima citados buscam fazer previsões a nível de cidade ou país, sem levar em conta diferenças que podem existir em granularidade espaço-temporal mais fina (por exemplo, por bairros). Compreender e prever como a doença vai se comportar em cada região de uma cidade é fundamental para a definição de políticas públicas, visto que a transmissão de uma doença não ocorre de forma uniforme numa cidade.

Este artigo propõe melhorias em modelos preditivos epidemiológicos a nível de bairro e consequentemente possam fazer previsões da cidade com maior precisão. Para isso, será utilizado um conjunto de dados georreferenciados contendo quatorze anos de casos de dengue na cidade de Fortaleza. A tarefa envolve prever, com base em dados das primeiras semanas do ano, o número de casos de dengue para o restante do ano em cada um dos 119 bairros da cidade, bem como para a cidade como um todo, identificando os bairros propensos a epidemia. A abordagem proposta envolve o uso de informações heurísticas que capturam a intensidade de propagação da doença em cada bairro, normalizada pela população do bairro, dentro de um modelo de rede neural baseado em LSTM. Avaliações quantitativas demonstram que o modelo prevê com maior precisão os casos de dengue em nível de bairro ao utilizar tais informações, evitando a tendência dos modelos baseados em auto-regressão de gerar previsões homogêneas para todos os bairros.

A estrutura do artigo é montada de forma sistemática em diversas etapas. Na Seção 2, uma visão geral do perfil temporal da dengue na cidade de Fortaleza é apresentada, enfatizando características de anos epidêmicos e não epidêmicos. Na Seção 3, é proposta uma heurística para formalizar a intensidade de propagação da doença como uma função da população local. Modelos de redes neurais baseados em LSTM são treinados com e sem utilizar a heurística proposta. Análises empíricas dos resultados e conclusões sobre os mesmos são apresentados na Seção 4 e Seção 5 respectivamente.

2 Dengue em Fortaleza

Fortaleza, a quinta maior capital do Brasil, é composta por 119 bairros e possui uma população total de $N = \sum_i^n N_i \approx 2.4 \times 10^6$, onde N_i representa a população de cada bairro i [14]. Para as análises neste artigo, foi utilizada uma base de dados de casos de dengue em Fortaleza. O Departamento de Saúde da cidade forneceu dados agregados e

anonimizados sobre os casos de dengue relatados por hospitais e postos de saúde pública. Os registros incluem o endereço residencial do paciente e a data de notificação, o que permite a agregação de dados nos 119 bairros de Fortaleza em uma base semanal. A série histórica completa pode ser acessada por meio do Sistema de Monitoramento Diário de Ocorrência de Doenças (SIMDA) [13].

A dengue é uma doença sazonal que depende da estação chuvosa para a proliferação do seu mosquito transmissor, o Aedes aegypti. A série temporal mostrada na Figura 1(a), que abrange o período de 2007 a 2020 e totaliza 222.817 casos, ilustra essa dependência. Tipicamente, a curva epidemiológica exibe uma forma de sino, com períodos de pico de contaminação ocorrendo entre os meses chuvosos ou imediatamente após as chuvas em abril e junho, variando em número total de casos a cada ano. Na Figura 1(a), é possível ver picos de casos mais intensos nos anos 2008, 2011, 2012, 2015, 2016 e 2017. É importante ressaltar a Figura 1(a) destaca o ano de 2020, quando restrições ao movimento das pessoas foram implementadas devido à pandemia de COVID-19. Houve uma tendência de aumento acentuado nos casos, o que poderia ter tornado-o em um ano epidêmico [9], mas isso não se concretizou, provavelmente devido ao Isolamento Social e o *Lockdown*. Destacamos que o período inicial que antecede as restrições de mobilidade apresentou casos totais de dengue similares aos anos de fortes epidemias 2008, 2011, 2012, 2015, 2016 e 2017.

Existe uma dispersão significativa na distribuição do total de casos de dengue entre os diferentes bairros de Fortaleza. Essa dispersão resulta em um número η de bairros epidêmicos para cada ano. Classificamos um bairro i como epidêmico quando o número total de casos acumulados até o final do ano ($C_T^{(i)}$) é maior do que um valor crítico C_{cri}. Bairros não epidêmicos são identificados quando $C_T^{(i)} < C_{cri}$. A Figura 1(b) exibe o valor de η para os vários anos analisados, com $C_{cri} = 100$ casos. A linha laranja com $\eta = 40$ separa os períodos analisados em dois conjuntos de anos denominados epidêmicos e não epidêmicos: $Y_{epi} = (2008, 2011, 2012, 2015, 2016, 2017)$ e $Y_{non-epi} = (2007, 2009, 2010, 2013, 2014, 2018, 2019, 2020)$, respectivamente. Em Y_{epi}, pelo menos 30% dos 119 bairros são considerados epidêmicos. Em $Y_{non-epi}$, são observados apenas pequenos surtos de casos de dengue em diferentes momentos do ano. Geralmente, o número de casos de dengue em cada bairro está relacionado à população local. A relação entre $C_T^{(i)}$ e N_i é mostrada na Figura 1(c). Tanto em Y_{epi} quanto em $Y_{non-epi}$, é observada uma correlação linear entre essas variáveis, com uma inclinação dependente do número total de casos registrados na cidade a cada ano. A relação representada na Figura 1(c) destaca o papel de N_i na quantificação de $C_T^{(i)}$ em diferentes anos. Assim, a população local influencia efetivamente a importância de cada bairro no número total de casos e regula a taxa de crescimento de casos per capita observada na cidade. Vale a pena enfatizar que variações em C_{cri} produzem valores de η diversos, porém não modifica a classificação dos conjuntos Y_{epi} e $Y_{non-epi}$.

Em geral, o comportamento da série temporal de cada bairro assemelha-se ao perfil observado na Figura 1(a). As divergências são principalmente notáveis em anos não epidêmicos. A propagação da dengue em Fortaleza ocorre de forma sincronizada durante os anos epidêmicos. Esse comportamento é evidente ao calcular o coeficiente de Pearson [10] para todas as combinações de bairros e para todos os anos no banco de dados. Os anos epidêmicos apresentam valores altos de coeficiente de Pearson, enquanto

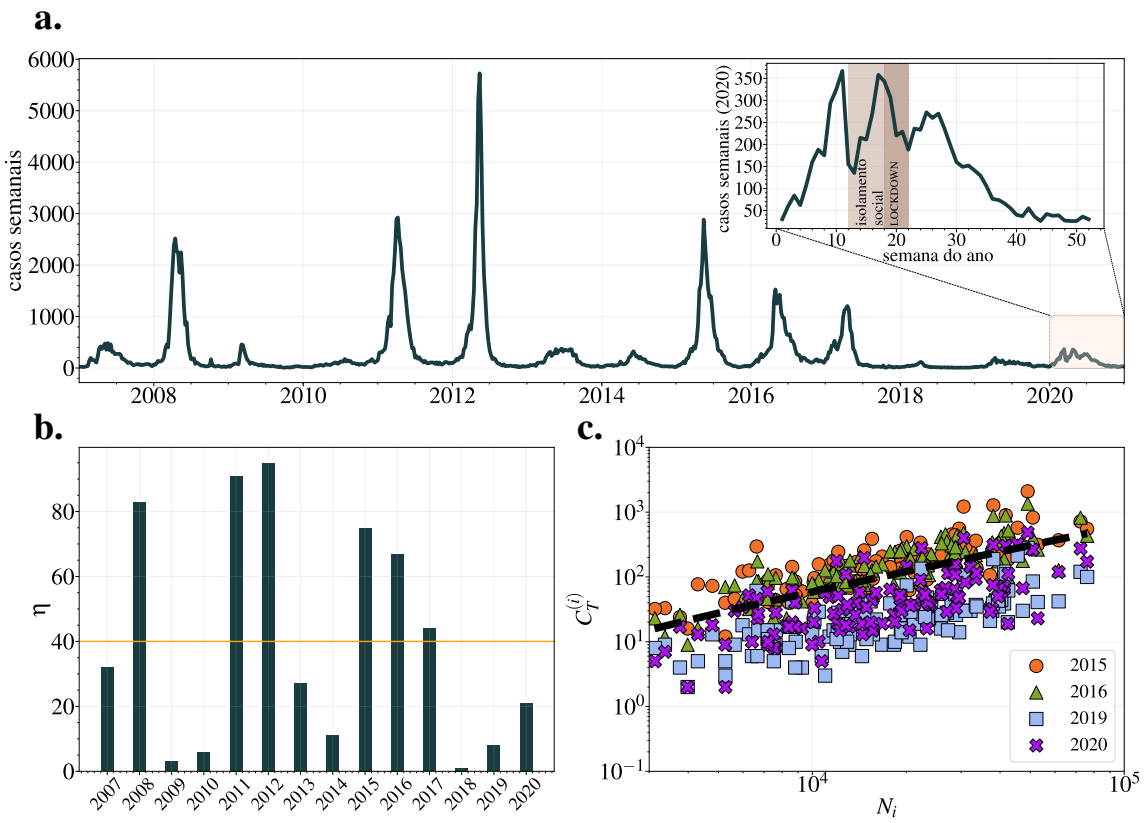

Figure 1. Visão geral dos casos de dengue em Fortaleza. A série temporal semanal de casos de dengue de 2007 a 2020 é apresentada em (a). O destaque mostra o período atípico de 2020, quando foram implementadas medidas de Isolamento Social e *Lockdown* para mitigar a propagação da COVID-19. Em (b), quantificamos o número de bairros epidêmicos, η, em cada período de estudo. Classificamos os bairros como epidêmicos quando o número total de casos observados em um determinado ano excede o valor de $C_{cri} = 100$ casos. A linha horizontal laranja separa os períodos analisados em anos epidêmicos ($Y_{epi} = (2008, 2011, 2012, 2015, 2016, 2017)$) e não epidêmicos ($Y_{non-epi} = (2007, 2009, 2010, 2013, 2014, 2019, 2020)$). Em (c), destacamos a relação entre o número total de casos observados no final de cada ano, $C_T^{(i)}$, em cada bairro i, em função da população N_i. A relação $C_T^{(i)} \times N_i$ apresenta um comportamento linear, destacando o papel da população local no número total de casos. A linha preta tracejada serve apenas como guia visual.

os anos não epidêmicos apresentam valores baixos. Acreditamos que erros potenciais de previsão podem ser mitigados por meio de um conjunto de dados de treinamento que considere os efeitos de correlação entre as séries temporais. Isso pode ser alcançado incorporando informações que levem em conta a população local como um fator relevante na taxa de crescimento dos casos. Na próxima seção, estabeleceremos uma conexão entre a intensidade de propagação da dengue ao longo do tempo e a população local.

3 Metodologia

3.1 A função de controle de intensidade (ICF)

A propagação da dengue em áreas urbanas pode ser particularmente desafiadora devido à complexidade do ambiente urbano e à interação entre os fatores que influenciam

a transmissão da doença. No entanto, tais fatores estão relacionados à população total do ambiente urbano. A população local torna-se uma variável relevante para controlar a intensidade da propagação da dengue, uma vez que a série temporal semanal de casos de dengue $W_c(t)$ pode ser integrada da seguinte forma:

$$C(t) = \sum_{t_0 \leq t} W_c(t_0), \tag{1}$$

onde $C(t)$ é a série temporal do número de casos acumulados na semana $t = 1, 2, 3, ...t_f$ e $C(t = t_f) = C_T$. Ressaltamos que existe uma relação linear em $C_T \times N$ (ver Figura 1(c)). A equação 1 gera um comportamento semelhante a uma função logística dada por [16]:

$$C(t) = \frac{L}{1 + exp(-k(t - t_{cri}))}, \tag{2}$$

que pode ser modificada para a seguinte forma:

$$\ln\left(\frac{C(t)}{L - C(t)}\right) = k(t - t_{cri}). \tag{3}$$

Considerando apenas o lado esquerdo da Equação 3, temos que:

$$h(L) = \ln\left(\frac{C(t)}{L - C(t)}\right) \tag{4}$$

onde L, k e t_{cri} são constantes. t_{cri} representa o momento efetivo de máximo de infecção semanal, com $C(t = t_{cri}) = C_T/2$. Para ajustar os dados à equação, é necessário conhecer a série temporal completa, tornando impraticável utilizá-la na previsão de casos de dengue nas primeiras semanas do ano. Em particular, a constante L governa a quantidade efetiva de casos totais observados no final do ano ($L \approx C_T \propto N$).

Na Figura 2, é mostrado o comportamento da Equação 4 para $L = C_T$ e $L = N$ para a cidade de Fortaleza nos anos de 2015, 2016, 2019 e 2020. Para cada ano, o momento t_{cri} é destacado através da igualdade $C = C_T/2$ (Equação 2). Observa-se que $h(L = C_T)$ e $h(L = N)$ diferem em escala para o mesmo número de casos observados. No entanto, para os períodos anteriores ao pico de contaminação ($t < t_{cri}$), as funções apresentam taxas de crescimento semelhantes ($\frac{dh}{dC}$). Para avaliar a hipótese de que as distribuições dos valores de $\frac{dh(L=C_T)}{dC}$ e $\frac{dh(L=N)}{dC}$ são iguais, foi realizado o teste de Kolmogorov-Smirnov [15]. O resultado do teste indica a distância máxima entre as duas FDA (função de distribuição acumulativa). Considerando um valor de significância de 0.05, as análises mostraram que a distância é de 0.43 ($p = 0.14$), indicando que as distribuições são estatisticamente iguais. Assim, decidimos definir a Função de Controle

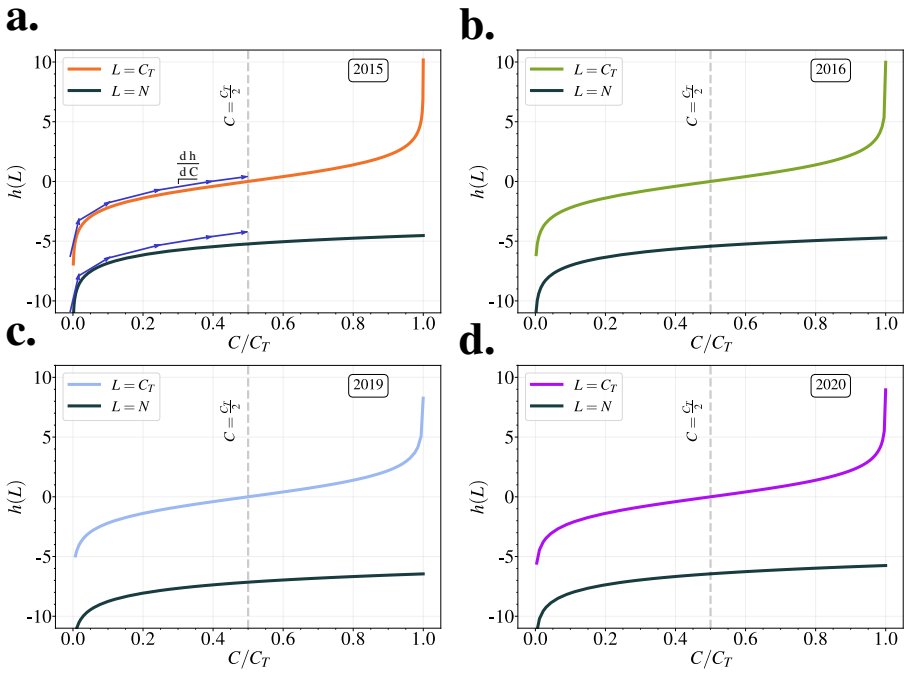

Figure 2. Comportamento da Equação 4 em função do número acumulado de casos ($C(t)$) para dois cenários: $L = C_T$ e $L = N$, para os anos de 2015, 2016, 2019 e 2020 (a-d, respectivamente). O eixo x foi normalizado em relação ao número total de casos observados ao final de cada período (C_T). Adicionalmente, uma linha vertical tracejada indica o momento de máximo de contaminação semanal ($C(t = t_{cri}) = C_T/2$). Antes de atingir o pico de infecção ($t < t_{cri}$), ambas as curvas $h(L = C_T)$ e $h(L = N)$ apresentam taxas de crescimento semelhantes. Observamos que $\frac{dh(L=C_T)}{dC} \approx \frac{dh(L=N)}{dC}$ (representado pelas setas azuis). Essa similaridade implica que a população local pode atuar como regulador para a taxa de crescimento dos casos.

de Intensidade (ICF, do inglês Intensity Control Function) da propagação da dengue como:

$$\begin{aligned} f(N, C(t)) &= \frac{dh(L = N)}{dC} \\ f(N, C(t)) &= \frac{N}{C(t)[N - C(t)]} \end{aligned} \tag{5}$$

A Equação 5 pode ser calculada para todos os bairros ao longo do período epidêmico. Valores altos de $f(N, C(t))$ são encontrados nos primeiros momentos de um surto de dengue, enquanto valores baixos de $f(N, C(t))$ são observados durante processos de grande infecção. Além disso, $f(N_1, C(t)) > f(N_2, C(t))$ quando $N_1 < N_2$. Dessa forma, $f(N, C(t))$ mitiga o processo de propagação da dengue, o que significa que valores menores em $f(N, C(t))$ regulam velocidades de propagação efetivas mais altas da doença em áreas urbanas. A hipótese é que a informação produzida por ICF contribui para uma melhor previsão da série temporal, uma vez que também captura o tamanho da população na taxa de crescimento dos casos.

3.2 Modelos neurais preditivos

Foram desenvolvidas duas arquiteturas de redes neurais baseadas em *Long short-term memory* (LSTM) para determinar se o ICF afeta a previsão de casos de dengue. LSTMs foram escolhidas para a previsão de casos de dengue por sua finalidade de aprendizado de dependências temporais longas, como é o caso da evolução de casos do vírus da dengue. Ambas as arquiteturas propostas têm como objetivo prever casos semanais de dengue em nível de bairro, especificamente para os 119 bairros de Fortaleza. A primeira arquitetura permite a entrada de dados na forma de uma matriz com dimensões 52×119, fornecendo à rede neural uma série temporal de tamanho 52 como entrada. Isso representa um histórico de 52 semanas para cada um dos 119 bairros em Fortaleza, denotados como ($X_{b1_{t51}}$, $X_{b1_{t50}}$, ... , X_{b1_t}, $X_{b2_{t51}}$, $X_{b2_{t50}}$, ... , X_{b2_t}, ..., $X_{bn_{t51}}$, $X_{bn_{t50}}$, ... , X_{bn_t}). A série temporal de entrada pode ser o histórico de casos de dengue, dados de incidência ou dados gerados pelo ICF. A saída da arquitetura é uma previsão de uma semana à frente para todos os bairros simultaneamente, representada como $(Y_{b1_{t+1}}, Y_{b2_{t+1}}, \ldots, Y_{bn_{t+1}})$ e as predições são realizadas de forma recursiva, a predição para o tempo $t + 1$ é utilizada também para prever valores no tempo $t + 2$. Identificamos uma elevada autocorrelação entre as séries temporais dos diversos anos para o intervalo de 52 semanas. Desta forma, foi selecionado um histórico de 52 semanas como entrada para a rede neural devido à natureza sazonal da dengue em Fortaleza.

A segunda arquitetura é semelhante à primeira, mas inclui uma série temporal transformada usando o ICF, além da série temporal de casos de dengue nos bairros. Assim, a entrada da segunda arquitetura é uma matriz com dimensões 52×238. A Figura 3 ilustra as duas arquiteturas de redes neurais, incluindo suas entradas e saídas.

Aplicamos quatro testes que utilizam dados de entrada e arquiteturas de redes neurais diferentes para prever os casos semanais de dengue em nível de bairro. A lista a seguir descreve as variações que serão comparadas:

- **RE** - teste realizado com a arquitetura representada na Figura 3(a) utilizando a série temporal de casos de dengue nos 119 distritos como entrada para a rede neural.
- **ICF** - teste realizado com a arquitetura representada na Figura 3(a), tendo como entrada para a rede neural a série temporal de casos de dengue nos 119 bairros modificada pela Equação 5.
- **ICF + RE** - teste realizado com a arquitetura representada na Figura 3(b) utilizando duas matrizes como entrada para a rede neural, uma contendo a série temporal de casos de dengue nos 119 bairros e outra contendo a série temporal modificada pela Equação 5.
- **INC** - teste realizado com a arquitetura representada na Figura 3(a) utilizando a série temporal de incidências de casos de dengue nos 119 distritos como entrada para a rede neural. Aqui, incidência ($w^{(i)}$) é definida pela relação:

$$w^{(i)}(t) = \frac{W^{(i)}(t)}{N_i} \tag{6}$$

onde $W^{(i)}$ é a série temporal semanal de casos semanais e N_i é a população do bairro i.

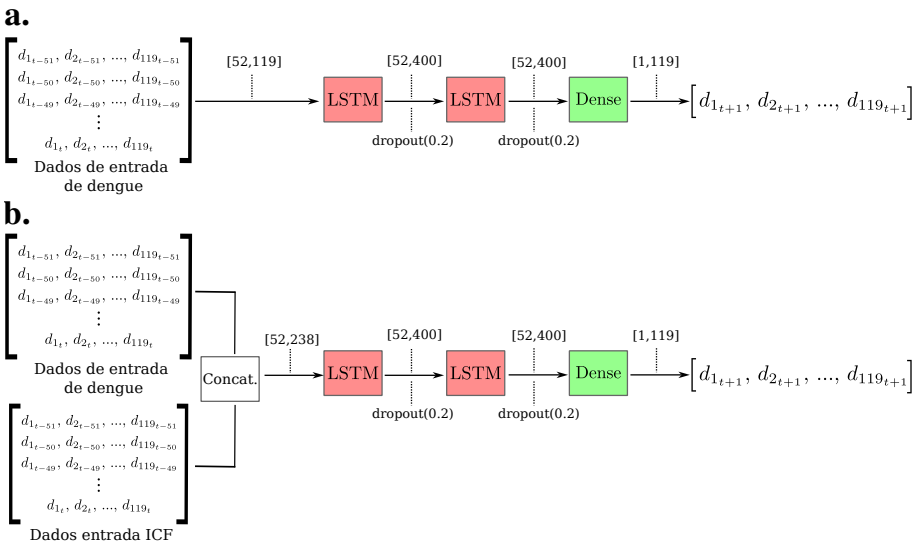

Figure 3. As duas redes neurais usadas neste artigo. Entre colchetes estão as dimensões dos dados passados entre cada camada da rede neural. O número de neurônios (400) da LSTM e o valor de dropout (0.2) apresentados na figura representam os valores ótimos encontrados pelo *Grid Search*. Além desses valores a rede neural foi treinada por épocas de 300 e *batch size* 64. Em (a) é apresentada a arquitetura da primeira rede neural, que recebe como entrada os casos de dengue ($W^{(i)}(t)$), a incidência de casos de dengue ($w^{(i)}(t)$) ou apenas os casos de dengue transformados por ICF (conforme mostrado na Equação 5). Em (b) é apresentada a arquitetura da segunda rede neural, que recebe como entrada tanto a série semanal de casos de dengue ($W^{(i)}(t)$) quanto suas modificações através do ICF.

3.3 Treinamento e teste

Para validar as variações dos testes e prever os casos de dengue para os 119 bairros, os modelos de rede neural foram implementados utilizando a biblioteca Keras [18] e foram treinados utilizando dados de 2007 a 2014. Os parâmetros da rede neural, incluindo épocas, método de otimização, número de neurônios, *dropout* e tamanho da pilha de LSTM, foram otimizados usando o método *Grid Search* fornecido pela biblioteca scikit-learn [17]. Esse método pesquisa por todos os valores pré-definidos para cada parâmetro a fim de identificar a combinação ideal. Para o processo de treinamento e validação os dados foram agrupados utilizando a técnica de *sliding window*.

Durante o processo de treinamento da rede neural, dados de dois anos foram separados para validação: um ano epidêmico e um ano não epidêmico. Para testar o desempenho dos modelos, foram utilizados dois anos epidêmicos (2015 e 2016) e dois anos não epidêmicos (2019 e 2020). A divisão dos dados para treinamento, validação e teste da rede neural para a previsão de casos de dengue nos quatro anos é apresentada na Tabela 1.

Sabendo que o pico de casos de dengue em Fortaleza ocorre em abril ou maio e que é de interesse público saber antecipadamente se ocorrerá ou não uma epidemia de casos de dengue, as previsões feitas nos testes foram realizadas a partir da semana 9, considerando que, para todos os 4 anos de teste, a semana 9 é antes do pico de casos de dengue.

Treinamento	Validação	Teste
2007, **2008**, **2011**, **2012**, 2013, 2014	2010, **2016**	**20015**
2007, **2008**, **2011**, **2012**, 2013, 2014	2010, **2015**	**2016**
2007, **2008**, **2011**, **2012**, 2013, 2014	2010, **2016**	2019
2007, **2008**, **2011**, **2012**, 2013, 2014	2010, **2016**	2020

Table 1. Separação de dados para treinamento, validação e teste das arquiteturas propostas de redes neurais. Anos em negrito são anos epidêmicos.

3.4 Métodos de avaliação

As previsões para os 119 bairros foram avaliadas com base em quatro objetivos: prever a intensidade máxima de casos, prever o tempo de pico, prever a série temporal completa de casos de dengue e classificar os anos e bairros como epidêmicos ou não epidêmicos.

Para avaliar as previsões de intensidade máxima, calculamos o erro médio absoluto logarítmico ($MALE_p$), entre os casos de dengue reais e previstos no pico:

$$MALE_p = \frac{\sum_{k=0}^{n} |log(y_k + 1) - log(\hat{y}_k + 1)|}{n}, \tag{7}$$

onde y é o número real de casos de dengue durante a semana de pico e $\hat{y}$ é o número previsto de casos de dengue durante a semana de pico.

O erro absoluto médio (MAE_t) representa a diferença entre os casos reais e previstos de dengue no pico, em termos da semana em que o pico ocorre:

$$MAE_t = \frac{\sum_{k=0}^{n} |r_k - \hat{r}_k|}{n}, \tag{8}$$

onde r é o índice da semana de pico da série temporal real, $\hat{r}$ é o índice da semana de pico da série temporal prevista, e n é o número total de bairros previstos.

O erro logarítmico médio-quadrático ($RMSLE$) mede a diferença entre as séries temporais reais e previstas, sendo calculada como:

$$RMSLE = \frac{1}{n} \sum_{i=1}^{n} \sqrt{\frac{\sum_{w=p}^{52} (log(x_w^{(i)} + 1) - log(\hat{x_w^{(i)}} + 1))^2}{52 - p}}, \tag{9}$$

onde x representa o número total de casos registrados de dengue, $\hat{x}$ representa o número total de casos de dengue previstos, n representa o número total de bairros, e p representa o tempo em que a previsão começa. Neste caso, $p = 9$, o que corresponde à semana do ano em que a previsão foi iniciada. Como a série histórica de casos de dengue em Fortaleza normalmente exibe apenas alguns picos por ano, o objetivo principal de um modelo de previsão é identificar o pico e seus altos e baixos. $RMSLE$ é a métrica mais apropriada para capturar isso, pois penaliza grandes diferenças entre os valores previstos e reais em períodos críticos, quando a incidência é alta.

Por fim, uma classificação de bairros entre epidêmicos ou não epidêmicos é medida usando a métrica $F1$-score. Note que bairros epidêmicos são aqueles que têm um número total de casos maior que C_{cri} no final do ano (semana 52).

4 Resultados

Dado o interesse do governo em identificar picos nos casos de dengue com antecedência e o fato de que esses picos geralmente ocorrem em torno das semanas $16 \leq \tau \leq 20$ do ano, é crucial que um modelo de rede neural que preveja casos de dengue para um ano completo possa prever com precisão os casos com algum aviso prévio para o período da semana τ. Portanto, em nossos testes, as previsões começam a partir da semana 9.

As Figuras 4(a)-(d) mostram os resultados das nossas previsões para os anos de 2015, 2016, 2019 e 2020, que foram obtidos utilizando os testes RE, $ICF + RE$, INC e ICF. As curvas ilustram o perfil temporal para a cidade de Fortaleza, onde a linha preta representa os registros reais de casos de dengue, e as outras linhas representam as previsões obtidas a partir dos diferentes testes.

Os primeiros surtos de dengue são observados nas primeiras semanas do ano, e os casos aumentam até períodos próximos a τ com o perfil de intensidade apropriado. Após o pico, os casos semanais diminuem com intensidade semelhante. No entanto, esse comportamento não é claramente capturado pela maioria dos testes, especialmente o teste INC. Os resultados de previsão do teste INC indicaram que os dados modulados pela incidência (conforme mostrado na Equação 6) não são relevantes o suficiente para que o modelo de rede neural aprenda a evolução dos casos de dengue. Uma situação semelhante é observada quando apenas os dados de ICF são utilizados. No entanto, vale ressaltar que a variação de ICF antecede o momento do pico de casos de dengue.

Quando analisamos apenas o modelo RE, observamos que para o ano de 2019 (um ano não epidêmico), o modelo foi capaz de acompanhar a curva de evolução dos casos reais de dengue. No entanto, o mesmo não ocorreu com os outros anos (2015, 2016 e 2020). Por outro lado, quando as informações de controle de intensidade são adicionadas, formando o teste $ICF + RE$, o modelo foi capaz de prever melhor a evolução dos casos de dengue, principalmente para os anos epidêmicos.

A Figura 5 mostra os resultados para as métricas detalhadas anteriormente. O painel indica que o teste $ICF + RE$ apresentou, efetivamente, um melhor desempenho na previsão da série temporal completa, no momento do pico e na intensidade do pico.

O teste RE apresentou resultados piores para $MALE_p$, MAE_t e $RMSLE$ em 2015 em comparação com todos os outros testes. No entanto, para 2016, os resultados foram semelhantes aos obtidos pelo teste INC. Já para 2019 e 2020, o teste RE apresentou melhores resultados quando comparado aos testes INC e ICF.

Embora a variação INC tenha mostrado bons resultados na previsão da intensidade do pico e do momento do pico para os anos de 2015, 2016 e 2020, esse teste apresentou resultados ruins para todas as variações em 2019, indicando grande instabilidade na previsão. Por fim, o teste RE mostrou resultados muito estáveis e teve um bom desempenho para todos os anos quando comparado aos testes ICF e INC.

A avaliação da qualidade das previsões por bairro foi feita a partir da correlação de Pearson entre as séries temporais previstas e reais no nível de bairro para os 4 anos em análise. Os resultados indicam que o teste com maior capacidade de prever a evolução temporal dos casos de dengue é o teste $IFC + RE$, seguido pelo teste RE. Por outro lado, os testes INC e ICF apresentaram os piores resultados.

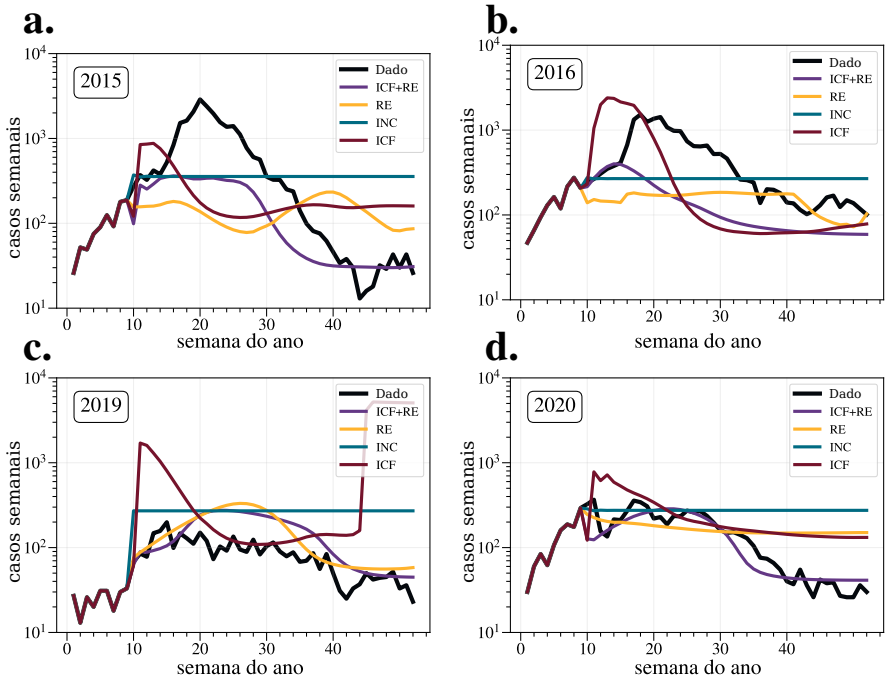

Figure 4. Séries temporais reais (Dado) e previstas para os 119 bairros acumulados semanalmente para os anos de 2015, 2016, 2019 e 2020, considerando as quatro variações de teste ($ICF + RE$, RE, INC e ICF)).

Os testes RE e $RE + ICF$ mostraram melhores resultados de previsão, contudo é importante identificar quais testes fornecem melhor desempenho na classificação de bairros epidêmicos. A Figura 6(a) exibe os valores de pontuação $F1$ para a classificação de bairros epidêmicos ou não epidêmicos para os testes $ICF + RE$ e RE. Os resultados mostram que o teste $ICF + RE$ apresentou melhor desempenho para todos os anos analisados.

Também investigamos quão genéricos são os testes $ICF + RE$ e RE, variando a semana inicial das previsões e o valor de C_{cri}. As pontuações $F1$ obtidas a partir desses testes para os anos de 2015, 2016, 2019 e 2020 foram comparadas (veja nas Figura 6(b)-(e)). Os resultados mostraram que o teste $ICF + RE$ teve um desempenho superior ao teste RE na previsão de casos de dengue para as semanas 9, 11 e 12 em 2015 e para todas as semanas iniciais em 2016. No entanto, para os anos não epidêmicos de 2019 e 2020, as pontuações $F1$ obtidas em ambos os testes foram semelhantes.

5 Conclusão

O artigo apresentou uma nova representação dos casos de dengue que permite que um modelo de rede neural aprenda a evolução da doença e extraia a relevância de cada bairro na transmissão total da cidade. Para isso, foi proposta uma transformação nos dados de casos de dengue com a hipótese de que isso limitaria a influência de cada bairro em relação à sua população residente.

O uso conjunto de informações sobre casos de dengue e dados gerados pela Função de Controle de Intensidade (ICF) levou a melhores resultados. Os testes que utilizaram formas isoladas de ICF ou INC não foram capazes de aprender a curva de

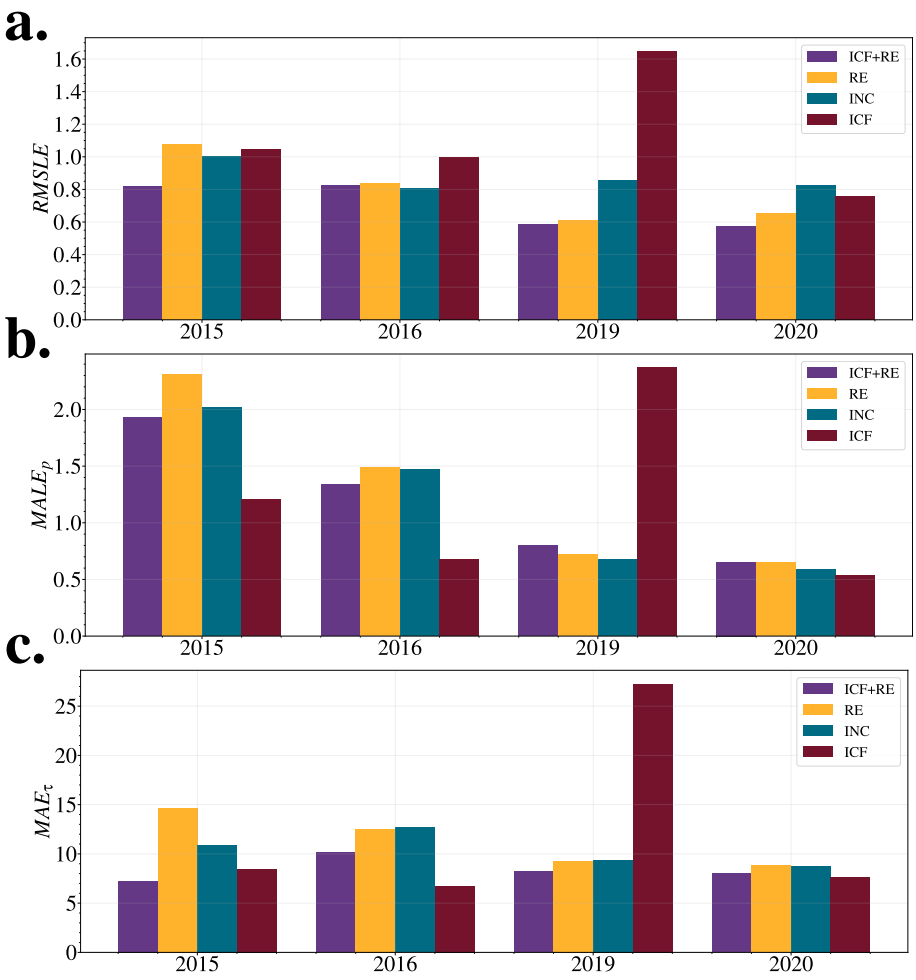

Figure 5. (a) $RMSLE$ entre os dados previstos e reais, (b) $MALE$ entre o valor previsto e real do pico de casos e (c) MAE entre a semana real e prevista do pico.

evolução dos casos nos anos testados.

De modo geral, o uso de uma função de controle de intensidade (ICF) traz benefícios para a previsão de casos de dengue em situações em que a rede neural precisa aprender a evolução da doença em várias localidades ao mesmo tempo e entender o impacto de cada localidade na previsão como um todo. Embora usar apenas casos de dengue para aprender a evolução dos casos seja uma estratégia competitiva, um modelo que considere apenas essa informação tende a prever picos de casos na mesma área e na mesma época do ano. O ICF incorpora informações limitando os casos e a população residente de cada bairro no modelo de rede neural, permitindo que ele considere a importância de cada bairro na transmissão da dengue para outros bairros e evite previsões de valores médios.

As melhorias descritas aqui fornecem evidências de que a heurística ICF pode ser utilizada em domínios relacionados. Trabalhos futuros devem buscar comparações com dados de outras doenças para demonstrar a generalidade do método apresentado.

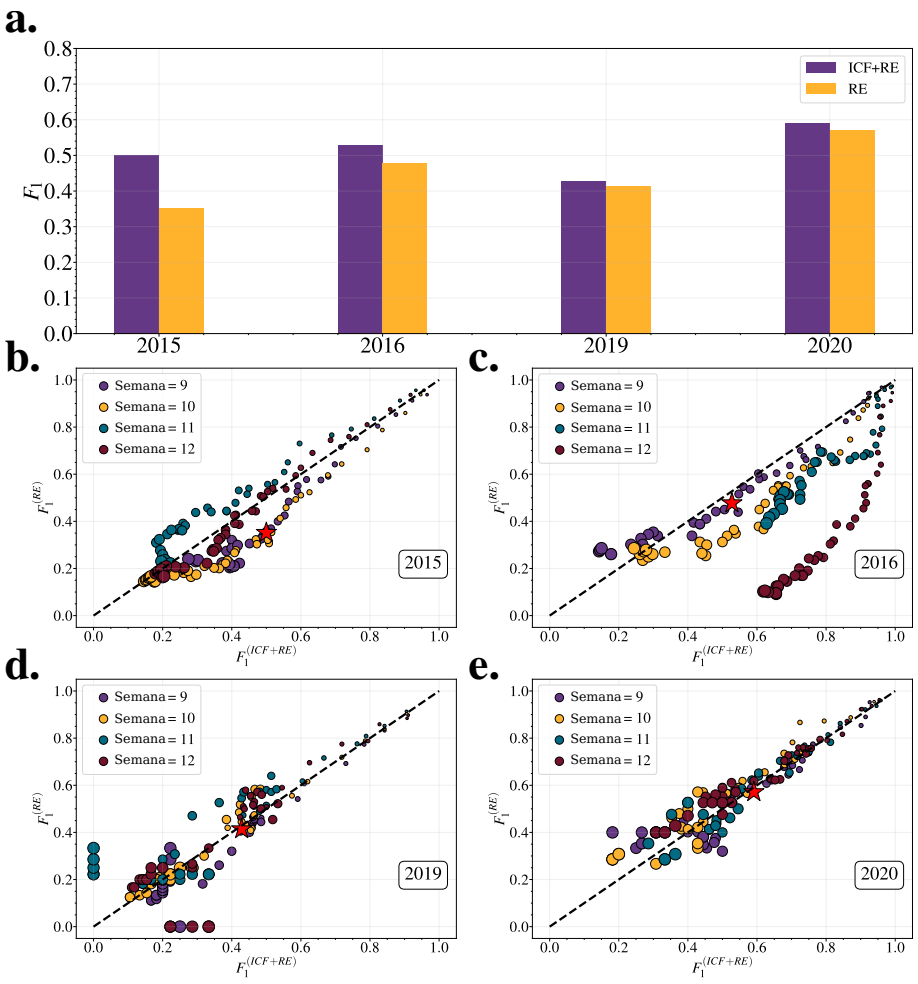

Figure 6. (a) valores de $F1$ dos testes RE e $ICF+RE$ para os anos de 2015, 2016, 2019 e 2020 com previsão a partir da semana 9. (b)-(e) mostram os valores $F_1^{(RE)}$ e $F_1^{(ICF+RE)}$, respectivamente, para os anos de 2015, 2016, 2019 e 2020. O eixo "Semana" representa a semana de início da previsão, que foi variada de 9 a 12. O tamanho de cada ponto corresponde ao valor de C_{cri} utilizado no experimento, com $C_{cri} \in [10, 200]$. A estrela vermelha indica o experimento realizado na semana 9 ($p = 9$) com $C_{cri} = 100$.

References

[1] Brockmann, D. & Helbing, D. The hidden geometry of complex, network-driven contagion phenomena. *Science*. **342**, 1337-1342 (2013)

[2] Araújo JLB, Oliveira EA, Lima Neto AS, Andrade JS Jr, Furtado V. Unveiling the paths of COVID-19 in a large city based on public transportation data. Sci Rep. 2023 Apr 8;13(1):5761. doi: 10.1038/s41598-023-32786-z. PMID: 37031258; PMCID: PMC10082688.

[3] Manrique-Saide, P., Coleman, P., McCall, P., Lenhart, A., Vázquez-Prokopec, G. & Davies, C. Multi-scale analysis of the associations among egg, larval and pupal surveys and the presence and abundance of adult female Aedes aegypti (Stegomyia aegypti) in the city of Merida, Mexico. *Medical And Veterinary Entomology*. **28**, 264-272 (2014)

[4] Pilger, D., Lenhart, A., Manrique-Saide, P., Siqueira, J., Da Rocha, W. & Kroeger, A. Is routine dengue vector surveillance in central Brazil able to accurately monitor the Aedes aegypti population? Results from a pupal productivity survey. *Tropical Medicine & International Health*. **16**, 1143-1150 (2011)

[5] Caminha, C., Furtado, V., Pequeno, T., Ponte, C., Melo, H., Oliveira, E. & Andrade Jr, J. Human mobility in large cities as a proxy for crime. *PloS One*. **12**, e0171609 (2017)

[6] Caminha, C., Furtado, V., Pinheiro, V. & Silva, C. Micro-interventions in urban transportation from pattern discovery on the flow of passengers and on the bus network. *2016 IEEE International Smart Cities Conference (ISC2)*. pp. 1-6 (2016)

[7] He, Z. & Tao, H. Epidemiology and ARIMA model of positive-rate of influenza viruses among children in Wuhan, China: A nine-year retrospective study. *International Journal Of Infectious Diseases*. **74** pp. 61-70 (2018)

[8] Kandula, S. & Shaman, J. Near-term forecasts of influenza-like illness: An evaluation of autoregressive time series approaches. *Epidemics*. **27** pp. 41-51 (2019)

[9] Bomfim, R., Pei, S., Shaman, J., Yamana, T., Makse, H., Andrade Jr, J., Lima Neto, A. & Furtado, V. Predicting dengue outbreaks at neighbourhood level using human mobility in urban areas. *Journal Of The Royal Society Interface*. **17**, 20200691 (2020)

[10] Schober, P., Boer, C. & Schwarte, L. Correlation coefficients: appropriate use and interpretation. *Anesthesia & Analgesia*. **126**, 1763-1768 (2018)

[11] Shahid, F., Zameer, A. & Muneeb, M. Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. *Chaos, Solitons & Fractals*. **140** pp. 110212 (2020)

[12] Kara, A. Multi-step influenza outbreak forecasting using deep LSTM network and genetic algorithm. *Expert Systems With Applications*. **180** pp. 115153 (2021)

[13] Sistema de Monitoramento Diário de Agravos (SIMDA). (http://tc1.sms.fortaleza.ce.gov.br/simda), Accessed: 2019-06-18

[14] Instituto Brasileiro de Geografia e Estatistica (IBGE). (http://www.ibge.gov.br), Accessed: 2019-06-18

[15] Conover, W. Practical nonparametric statistics. (john wiley & sons,1999)

[16] Verhulst, P. F. (1838). Notice sur la loi que la population poursuit dans son accroissement. Correspondance mathématique et physique, **10**: p. 113-121.

[17] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal Of Machine Learning Research*. **12** pp. 2825-2830 (2011)

[18] Chollet, F. & Others Keras. (https://keras.io,2015)