Klasifikasi Situs *Phishing* dengan Menggunakan Neural Network dan K-Nearest Neighbor

Slamet Widodo [1,*]

[1] Manajemen Informatika; AMIK BSI Pontianak; Jln. Abdurrahman Saleh Pontianak No.18A, telp/fax: 0561-583924, e-mail: dodo.swd@gmail.com

* Korespondensi: e-mail: dodo.swd@gmail.com

Diterima: 16 Mei 2017; Review: 23 Mei 2017; Disetujui: 30 Mei 2017

Cara sitasi: Widodo S. 2017. Klasifikasi Situs Phishing dengan Menggunakan Neural Network dan K-Nearest Neighbor. Information Management For Educators And Professionals. 1 (2): 145-154.

Abstrak: Meningkatnya jumlah pengguna internet dan toko *online* disertai maraknya jumlah situs *phishing*. Menurut laporan dari APWG, jumlah laporan *phishing* yang disampaikan selama kuartal 2 2016 adalah 466.065. Laporan *phishing* mengalami peningkatan dari 61% pada 289.371 yang diterima pada kuartal pertama 2016. Biasanya, serangan *phishing* dimulai dari sebuah *e-mail* yang tampaknya dikirim dari sebuah organisasi resmi kepada korbannya untuk memperbarui atau memvalidasi informasi terbaru mereka dengan mengikuti *link* URL dalam *e-mail* tersebut. Dengan cara ini awal serangan *phishing* dimulai dengan mengunjungi *link* yang diterima dalam sebuah e-mail. Hal menyebabkan kerugian ekonomis yang cukup signifikan, terutama bagi perusahaan. Hal-hal tersebut mendasari dilakukannya penelitian tentang klasifikasi situs *phishing* yang kemudian akan diklasifikasikan berdasar dua kategori utama yaitu situs *phishing* dan non-*phishing*. Pengklasifikasian pada penelitian ini diselesaikan dengan menggunakan metode NN dan KNN. NN banyak diterapkan dalam penelitian karena kemampuanya dalam memodelkan sistem yang sangat nonlinier di mana hubungan antara variabel-variabel tidak diketahui (generalisasi) atau sangat kompleks, sedangkan KNN atau *K- Nearest Neighbor* metode pengklasifikasian yang berdasar pada pengukuran jarak tetangga terdekat serta memiliki performansi yang baik ketika data training yang diberikan sedikit. *Neural network* dengan *backpropagation* mampu memberikan hasil klasifikasi sebesar 91.21% lebih besar dibanding dengan menggunakan metode *k-nn* dengan ketepatan klasifikasi 90.33%.

Kata kunci: *neural network, k-nearest neighbor*, situs, *phishing,*

Abstract: *The increasing number of Internet users and online shops along with the rise of the number of phishing sites. According to reports from the APWG, the number of phishing reports submitted during the second quarter of 2016 was 466 065. Reports of phishing has increased from 61% at 289 371 received in the first quarter of 2016. Typically, phishing attacks initiated from an e-mail that appears to be sent from a legitimate organization to the victim to update or validate them with the latest information follow the URL link in an e-mail the. In this way the beginning of a phishing attack starts by visiting the link received in an e-mail. It causes significant economic losses, especially for companies. These things underlie research about the classification of phishing sites which will then be classified by two main categories, namely non-phishing sites and phishing. The classification in this study resolved by using NN and KNN. NN widely applied in research because of his ability to model highly nonlinear systems in which the relationship between the variables is unknown (generalization) or very complex, while KNN or K-Nearest Neighbor classification method based on the nearest neighbor distance measurements and has a good performance when given a little training data. Neural networks with backpropagation able to provide the classification of 91.21% larger than using the k-nn with a classification accuracy of 90.33%.*

Keywords*: neural networks, k-nearest neighbor, websites, phishing.*

E-ISSN: 2548-3331 ; 145 – 154

1. Pendahuluan

Jumlah pengguna internet dalam transaksi belanja *online* meningkat dengan signifikan sebagai alternatif praktis dalam melakukan transaksi jual-beli. Penjualan di seluruh dunia dari toko *online* meningkat 20,1% pada tahun 2014 mencapai $1500 triliun (Ingham, Cadieux, & Berrada, 2014). Oleh karena itu, tujuan utama dari toko *online* yaitu untuk menarik sebanyak mungkin konsumen. Hal ini menyebabkan meningkatnya persaingan di antara toko *online* untuk memperkenalkan layanan berkualitas bagi konsumen. Namun meningkatnya jumlah toko *online* dan website disertai maraknya jumlah situs *phishing*. Meskipun pengguna internet menyadari telah banyak jatuh korban dari serangan *phishing* tersebut. Tujuan dari serangan ini adalah untuk membuat pengguna internet percaya bahwa mereka berinteraksi dengan situs *online* resmi. Situs *phishing* dapat muncul menjadi jenis website, termasuk situs pembayaran atau situs lelang *online*. Metode yang efisien untuk mengidentifikasi situs *phishing* diperlukan untuk melindungi dari data sensitif pengguna. Informasi yang dicari oleh *phisher* (pelaku *phishing*) adalah berupa *password*, akun atau nomor kartu kredit korban dengan cara mengirim *email, banner* atau *pop-up window* untuk menjebak pengguna mengarah ke situs web palsu dimana korban diminta untuk memberikan informsi pribadinya.

Selain berbahaya bagi pelanggan, serangan *phishing* juga merusak reputasi lembaga keuangan yang bersangkutan, karena pelanggan menjadi kurang percaya bahwa akun mereka dapat diakses dengan aman. situs *phishing* dianggap salah satu kejahatan elektronik yang paling umum (Aaron & Manning, 2016). Menurut laporan dari APWG, jumlah laporan *phishing* yang disampaikan selama kuartal 2 2016 adalah 466.065. Laporan *phishing* mengalami peningkatan dari 61% pada 289.371 yang diterima pada kuartal pertama 2016 (Aaron & Manning, 2016). Biasanya, serangan *phishing* dimulai dari sebuah *e-mail* yang tampaknya dikirim dari sebuah organisasi resmi kepada korbannya untuk memperbarui atau memvalidasi informasi terbaru mereka dengan mengikuti *link* URL dalam *e-mail* tersebut. Dengan cara ini awal serangan *phishing* dimulai dengan mengunjungi *link* yang diterima dalam sebuah *e-mail* . Pentingnya dilakukan penelitian ini dikarenakan penanganan situs yang efektif tidak hanya dapat mengurangi kerugian perusahaan tetapi juga meningkatkan kepuasan dari pengguna situs itu sendiri. Penelitian ini mengenai pendeteksian situs *phishing* akan digali dari fitur suatu situs, dengan cara menganalisis fitur yang membedakan situs *phishing* dengan situs non *phishing*. Dari fitur yang diberikan oleh analisis tersebut kemudian akan dilihat bagaimana karakteristik dari suatu situs *phishing* pada dataset yang digunakan. Kemudian hasil analisa akan di klasifikasi dengan menggunakan metode NN atau *Neuron Network* dan KNN atau *K-Nearest Neighbor*.

NN banyak diterapkan dalam penelitian karena kemampuanya dalam memodelkan sistem yang sangat nonlinier di mana hubungan antara variabel-variabel tidak diketahui (generalisasi) atau sangat kompleks (Amato, Lopez, Pena-Mendez, Vanhara, & Hampl, 2013). Kemampuan dari ANN telah dibuktikan pada beberapa aplikasi termasuk *speech synthesis*, diagnosa, bidang pengobatan, keuangan dan bisnis, kontrol pada robot, pemrosesan sinyal, dan masalah lain yang termasuk dalam kategori pengenalan pola dan klasifikasi (Jadav & Panchal, 2012). Teknik yang populer digunakan pada metode ANN adalah algoritma *BackPropagation* (*BP*). Sedangkan KNN atau *K- Nearest Neighbor* metode pengklasifikasian yang berdasar pada pengukuran jarak tetangga terdekat serta memiliki performansi yang baik ketika data training yang diberikan sedikit (Colas & Brazdil, 2006). Dalam pengenalan pola, K-NN adalah metode non-parametrik yang digunakan untuk klasifikasi dan regresi. pada kedua kasus tersebut, input terdiri dari K terdekat pada data pelatihan dalam ruang fitur. Dalam konteks mesin pembelajaran, metode ini telah banyak digunakan diberbagai masalah pembelajaran terawasi. Selain kemudahan dan ketidakcukupan teori yang menjamin untuk bisa digunakan untuk data set yang sedikit, kemampuan metode ini telah diamati dengan hasil yang sangat baik, bahkan dengan hasil diluar dugaan para pengklasifikasi (Piro, Nock, Nielsen, & Barlaud, 2012).

Dengan kelebihan dua algoritma inilah yang mendasari dilakukannya penelitian tentang klasifikasi situs *phishing* menggunakan metode *neural network* (NN) dan k–*nearest neighbour* (KNN) sehingga dapat diketahui hasil performansi yang diberikan oleh kedua metode tersebut.

2. Metode Penelitian

2.1 Desain Penelitian

Penelitian merupakan suatu proses mencari secara sistematis dalam waktu yang relatif lama dengan menggunakan aturan-aturan yang ada. Penelitian merupakan aktivitas dalam mempertimbangkan, dimana bertujuan untuk membuat konstribusi asli dalam pengetahuan (Dawson, 2007). Dalam konteks sebuah penelitian, pendekatan metode yang digunakan untuk memecahkan masalah, diantaranya: mengumpulkan data, merumuskan hipotesis atau proposisi, menguji hipotesis, hasil penafsiran, dan kesimpulan yang dapat dievaluasi secara independen oleh orang lain (Berndtsson, Hansson, & Lundell, 2008). Sedangkan menurut (Dawson, 2007) terdapat empat metode penelitian yang umum digunakan, diantaranya:

1. *Action Research* yaitu penerapan penelitian yang dilakukan dengan berfokus pada tindakan sosial.
2. *Experiment* yaitu penelitian yang dilakukan dengan cara dimana dapat dilakukan di laboratorium ataupun di lingkungan sebenarnya dengan cara simulasi.
3. *Case Study* yaitu penelitian yang bersifat penekanan dalam kasus-kasus yang spesifik.
4. *Survey* yaitu penelitian yang dilakukan dengan melakukan pengajuan pertanyaan tertulis atau yang dikenal dengan kuisioner ataupun dengan cara wawancara.

Penelitian ini menggunakan penelitian eksperimen. Penelitian eksperimen melibatkan penyelidikan perlakuan pada parameter atau variabel tergantung dari penelitinya dan menggunakan tes yang dikendalikan oleh si peneliti itu sendiri, dengan metode penelitian sebagai berikut:

1. Pengumpulan data
   Pada tahap ini ditentukan data yang akan diproses. Mencari data yang tersedia, memperoleh data tambahan yang dibutuhkan, mengintegrasikan semua datakedalam data set, termasuk variabel yang diperlukan dalam proses.
2. Pengolahan data awal
   Ditahap ini dilakukan penyeleksian data, data dibersihkan dan ditransformasikan kebentuk yang diinginkan sehingga dapat dilakukan persiapan dalam pembuatan model.
3. Metode yang diusulkan
   Pada tahap ini data dianalisis, dikelompokan variabel mana yang berhubungan dengan satu sama lainnya. Setelah data dianalisis lalu diterapkan model-model yang sesuai dengan jenis data. Pembagian data kedalam data latihan (*training data*) dan data uji (*testing data*) juga diperlukan untuk pembuatan model.
4. Eksperimen dan pengujian metode
   Pada tahap ini model yang diusulkan akan diuji untuk melihat hasil berupa *rule* yang akan dimanfaatkan dalam pengambilan keputusan.
5. Evaluasi dan validasi
   Pada tahap ini dilakukan evaluasi terhadap model yang ditetapkan untuk mengetahui tingkat keakurasian model.

2.2 Pengumpulan Data

Dalam pengumpulan data terdapat sumber data, sumber data yang terhimpun langsung oleh peneliti disebut dengan sumber primer, sedangkan apabila melalui tangan kedua disebut sumber sekunder (Riduwan, 2008). Data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh dari uci repository. Data *phishing* dengan jumlah data sebanyak 11055 situs sebanyak 6157 yang terdeteksi situs non-*phishing*, sedangkan 4898 situs terdeksi situs *phishing*. Data situs *phishing* terdri dari 18 variabel atau atribut. Variabel tersebut ada yang tergolong variabel pemrediksi yaitu variabel yang dijadikan sebagai penentu situs *phishing*, dan variabel tujuan yaitu variabel yang dijadikan sebagai hasil. Adapun variabel pemrediksi yaitu: *using_ip_address, long_URL, URL_having_@_symbol, adding_prefix_and_suffix, sub-domain, misuse_of_https, request_url, URL_of_anchor, server_from_handler, abnormal_URL, redirect_page, using_pop-up_window, dns_record, website_traffic, age_of_domain, disabling_right_click.* Untuk meningkatkan akurasi dan efisiensi

E-ISSN: 2548-3331 ; 145 – 154

algoritma. Data yang digunakan dalam penulisan ini bernilai kategorikal. Data ditransformasikan kedalam aplikasi Rapidminer. Tabel kategorikal atribut terlihat pada tabel 1.

Tabel 1. Tabel Transformasi Atribut

Nilai	**Atribut**								
1 = valid 0 = mencurigakan -1 = *phishing*	Using IP address	Long URL	URL having @ Symbol	Adding Prefix and Suffix	Sub-Domain(s)	Misuse of HTTPs	Request URL	URL of Anchor	Server Form Handler
	Abnormal URL	Redirect Page	Using Pop-up Window	Hiding Suspicious *Link*	DNS record	Website Traffic	Age of Domain	Disabling Right Click	

Sumber: Mohammad, Thabtah, & McCluskey (2014)

Dari data tabel diatas, 17 atribut digunakan untuk membedakan situs *phishing* dengan situs non-*phishing*, atribut tersebut dijelaskan sebagai berikut (Mohammad, Thabtah, & McCluskey, 2014):

1. *Using the IP Address*
   Menggunakan alamat IP di bagian hostname dari alamat URL dapat diindikasikan situs ini mencoba untuk mencuri informasi pribadi seseorang. Sebagai contoh alamat situs berupa "http://125.98.3.123/fake.html" atau kadangkala alamat situs ditransformasi menjad kode hexadesimal seperti "http://0x58.0xCC.0xCA.0x62/2/paypal.ca/index.html".

2. *Long URL*
   Situs *phishing* menggunakan URL yang panjang untuk menyembunyikan bagian yang mencurigakan pada address bar. Contohnya: http://federmacedoadv.com.br/3f/aze/ab51e2e319e51502f416dbe46b773a5e/?cmd=_home&dispatch=11004d58f5b74f8dc1e7c2e8dd4105e811004d58f5b74f8dc1e7c2e8dd4105e8@*phishing*.website.html. Dalam penelitiannya (Mohammad, Thabtah, & McCluskey, 2014) menjelaskan jika panjang karakter kurang dari URL 54 karaketr diklasifikasikan dalam situs valid, URL dengan karakter 54 sampai 75 karakter URL diklasifikasikan mencurigakan, sedangkan lebih 75 karakter maka URL diklasifikasikan *phishing*.

3. *URL having @ Symbol*
   Menggunakan simbol "@" pada alamat URL mengarahkan browser untuk mengabaikan segala sesuatu yang mendahului simbol "@". Jika suatu situs menggunakan simbol ini dikategorika kedalam *phishing*.

4. *Adding Prefix and Suffix*
   Tanda pisah (*dash symbol*) jarang digunakan dalam URL yang sah. Phisher (pelaku *phishing*) cenderung menambah awalan atau akhiran yang dipisahkan oleh tanda (-) untuk nama domain sehingga pengguna merasa bahwa mereka mengakses halaman web yang sah. Misalnya http://www.Confirme-paypal.com/.

5. *Sub-Domain(s)*
   Teknik lain yang digunakan oleh phisher (pelaku *phishing*) untuk menipu pengguna adalah dengan menambahkan sub-domain (s) ke URL sehingga pengguna dapat percaya bahwa mereka mengakses situs resmi. Contoh domain situs phising adalah : https://facebook.login.com (palsu), jika terlihat demikian maka itu adalah website phising. Karena domainnya adalah login.com bukan facebook.com. Jika jumlah titik (dot) lebih besar dari satu, maka URL tersebut diklasifikasikan sebagai "Mencurigakan" karena memiliki satu sub domain. Namun, jika titik-titik yang lebih besar dari dua, itu diklasifikasikan sebagai "*Phishing*" karena akan memiliki beberapa sub domain. Jika URL tidak memiliki sub domain, diklasifikasikan sebagai situs "sah".

6. *Misuse of HTTPs*
   Https dapat menjamin keamanan dalam Autentikasi server yaitu memungkinkan peramban dan pengguna memiliki kepercayaan bahwa mereka sedang berbicara kepada server aplikasi sesungguhnya. Https juga mampu dalam menjaga kerahasiaan data dan Integritas data. Phisher dapat menggunakan protokol HTTPS palsu sehingga pengguna mungkin saja bisa tertipu. Direkomendasi untuk memeriksa apakah protokol HTTPS disediakan oleh penerbit ternama seperti "GeoTrust, GoDaddy".

7. *Request URL*
   Sebuah halaman web biasanya terdiri dari teks dan beberapa objek seperti gambar dan video. Biasanya, objek-objek ini dimuat ke halaman web dari domain yang sama di mana halaman web yang ada. Jika objek yang diambil dari domain yang berbeda dari domain yang diletik di alamat URL maka halaman web adalah berpotensi mencurigakan.

8. *URL of Anchor*
   Mirip dengan "*Request URL"* tapi untuk fitur ini *link* dalam halaman web mungkin merujuk pada domain yang berbeda dari domain yang diketik pada address bar URL. Fitur ini diberlakukan persis seperti "*Request URL".*

9. *Server Form Handler*
   SFHs yang mengandung string kosong atau "about: blank" dianggap meragukan karena akan mengambil informasi yang disampaikan. Selain itu, jika nama domain di SFHs berbeda dari nama domain dari halaman web, ini mengungkapkan bahwa halaman web adalah curiga karena informasi yang disampaikan jarang ditangani oleh domain eksternal.

10. *Abnormal URL*
    Jika identitas situs tidak cocok riwayatnya dengan yang ditampilkan dalam database WHOIS (http://who.is/) website ini diklasifikasikan sebagai "*Phishing*".

11. *Redirect Page*
    Fitur ini biasanya digunakan oleh phisher dengan menyembunyikan *link* sebenarnya yang meminta pengguna untuk mengirimkan informasi mereka ke situs web yang mencurigakan.

12. *Using Pop-up Window*
    Hal ini yang tidak wajar jika situs yang sah meminta pengguna untuk mengirimkan identitasnya melalui jendela popup. Jika hal ini terjadi maka diklasifikasikan ke "*phishing*".

13. *Hiding Suspicious Link*
    Phisher akan menyembunyikan *link* yang mencurigakan dengan menunjukkan *link* palsu pada status bar browser atau dengan menyembunyikan status bar itu sendiri.

14. *DNS record*
    Jika catatan DNS kosong atau tidak ditemukan maka website ini diklasifikasikan sebagai "*Phishing*", jika tidak maka diklasifikasikan sebagai "sah". Rekaman DNS menyediakan informasi tentang domain yang masih hidup, sedangkan domain dihapus tidak tersedia pada catatan DNS.

15. *Website Traffic*
    Fitur ini mengukur popularitas website dengan menentukan jumlah pengunjung dan jumlah halaman yang mereka kunjungi. Situs *phishing* tidak ditemukan dalam database Alexa. Situs yang sah memiliki rentang 100.000 peringkat. Selain itu, jika domain tidak memiliki lalu lintas atau tidak diakui oleh database Alexa, diklasifikasikan sebagai "*Phishing*". Jika bukan keduanya diklasifikasikan sebagai "Mencurigakan".

16. *Age of Domain*
    Situs web dianggap "sah" jika usia domain lebih dari 2 tahun.

17. *Disabling Right Click*

E-ISSN: 2548-3331 ; 145 – 154

Phisher menggunakan JavaScript untuk menonaktifkan fungsi klik kanan, sehingga pengguna tidak dapat melihat dan menyimpan kode sumber halaman web. Jika fungsi klik kanan dinonaktifkan diklasifikasikan ke dalam '*phishing*".

2.3 Pengolahan Data Awal

Pengolahan awal data berupa pembentukan sumber data random (*set the random seed*) dan pembentukan variabel pemilihan (*partition variabel*). Untuk mendapatkan data yang dapat tersusun dengan baik maka diperlukan pembentukan sumber data random karena sudah mengalami pengacakan data secara statistik. Data yang digunakan berjumlah 11.055 record, dibagi menjadi dua untuk data training (80%) dan data testing (20%), dari 11.055 record akan digunakan untuk data training sebanyak 80% atau 8.844 record dan 20% atau 2.211 record digunakan sebagai data testing. Pengambilan data untuk data testing menggunakan teknik Systematic Random Sampling. Teknik systematic random sampling merupakan modifikasi dari teknik random sampling, dengan cara memilih subjek dari daftar populasi secara sistematis bukan secara acak. Pemilihan secara acak hanya digunakan untuk memilih data pertama saja. Penentuan data berikutnya dengan cara memanfaatkan interval sampel, yaitu angka yang menunjukan jarak antara nomor- nomor urut yang terdapat dalam kerangka sampling yang dijadikan sebagai patokan dalam menentukan data kedua dan seterusnya hingga data ke-n (Cohen, Manion, & Morrison, 2007). Untuk menentukan interval sampel menggunakan rumus berikut:

f=N/sn

dimana:

f = frekuensi interval

N = Jumlah total populasi

sn = Jumlah sample yang diperlukan

Total data *testing* yang diperlukan sebanyak 20% dari 11.055 record yaitu 2.211 record, dengan menggunakan rumus 2.1 untuk mendapatkan interval maka dari 11.055 record dibagi dengan 2.211 record sehingga didapat intervalnya adalah 5. Sehingga untuk pemilihan data berikutnya didasarkan pada nomor kelipatan lima.

2.4 Langkah Analisis

Langkah analisis yang dilakukan dalam penelitian ini adalah sebagai berikut:

1. Menyiapkan dan mengumpulkan data untuk eksperimen
2. Preprocessing data.
3. Mendapakatkan hasil *phishing* yang sudah di preprocessing
4. Melakukan pengacakan data dengan menggunakan 10-*fold cross validation*
5. Menganalisis karakteristik dari situs *phishing* dan non-*phishing*
6. Mendesain arsitektur neural network, dengan memasukkan nilai parameter *neural network* yaitu (*training cycles, learning rate, momentum, dan neuron size pada hidden layer*)
7. Klasifikasi data menggunakan metode *neural network* dengan *backpropagation.*
8. Klasifikasi data menggunakan metode *k-nearest neighbor* dengan k = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 dan membandingkan hasil klasifikasi yang diberikan oleh k yang berbeda pada KNN
9. Melakukan uji pada 11.055 data testing dengan metode yang memberikan hasil klasifikasi terbaik dan mengambil analisis.
10. Membandingkan performansi metode NN dan KNN berdasarkan tingkat ketepatan klasifikasi, *precision* dan *recall*.

3. Hasil dan Pembahasan

3.1. Hasil *Training* dengan *Neural Network*

Backpropagation merupakan salah satu metode pelatihan terbimbing (*supervised learning*) pada pelatihan *neural network*. Pada pelatihan terbimbing diperlukan sejumlah masukan dan target yang berfungsi untuk melatih jaringan hingga diperoleh bobot yang diinginkan. Setiap kali pelatihan, suatu input diberikan ke dalam jaringan. Jaringan akan memproses dan mengeluarkan keluaran. Selisih antara keluaran jaringan dengan target merupakan kesalahan yang terjadi. Jaringan akan memodifikasi bobot sesuai dengan kesalahan tersebut. Penelitian ini melakukan uji coba untuk mencari nilai *training cycles* dengan cara memasukkan nilai dengan *range* dari 100 sampai dengan 1000 *untuk training*

cycles, sedangkan untuk nilai *learning rate* 0.3 dan *momentum* 0.2 nilainya tidak diubah. Berikut ini adalah hasil dari percobaan yang telah dilakukan untuk penentuan nilai *training cycles*:

Tabel 2. Penelitian Penentuan Nilai *Training Cycles*

Training Cycles	Learning Rate	Momentum	Accuracy	AUC
100	0.3	0.2	90.86	0.968
200	0.3	0.2	90.99	0.968
Training Cycles	**Learning Rate**	**Momentum**	**Accuracy**	**AUC**
300	0.3	0.2	91.12	0.969
400	0.3	0.2	91.13	0.969
500	0.3	0.2	91.04	0.969
600	0.3	0.2	91.05	0.969
700	0.3	0.2	90.99	0.969
800	0.3	0.2	91.04	0.969
900	0.3	0.2	91.14	0.969
1000	0.3	0.2	91.21	0.969

Sumber: Hasil Penelitian (2016)

Dari hasil uji coba yang dilakukan pada nilai training *cycles* dipilih berdasarkan nilai akurasi dan nilai *Area Under Curve* (AUC) terbesar, dengan nilai 1000 pada training *cycles*. Kemudian nilai tersebut akan digunakan untuk percobaan selanjutnya yaitu untuk menentukan nilai Learning Rate. Nilai *Learning rate* ditentukan dengan cara melakukan uji coba dengan nilai range 0.1 sampai dengan 1.0. Untuk nilai momentum tetap 0.2 sedangkan nilai Training *cycles* dipilih dari percobaan sebelumnya yaitu 1000. Berikut adalah hasil dari eksperimen yang telah dilakukan untuk menentukan nilai Learning Rate:

Tabel 3. Penelitian Penentuan Nilai *Learning Rate*

Training *Cycles*	Learning Rate	Momentum	Accuracy	AUC
1000	0.1	0.2	90.76	0.969
1000	0.2	0.2	90.8	0.969
1000	0.3	0.2	91.21	0.969
1000	0.4	0.2	91.04	0.969
1000	0.5	0.2	91.16	0.969
1000	0.6	0.2	90.8	0.969
1000	0.7	0.2	90.89	0.969
1000	0.8	0.2	91.06	0.968
1000	0.9	0.2	90.75	0.968
1000	1	0.2	91.15	0.969

Sumber: Hasil Penelitian (2016)

Berdasarkan hasil uji coba pada nilai Learning Rate, nilai akurasi dan nilai Area *Under Curve* (AUC) terbesar yang dihasilkan 0.3. Nilai tersebut akan digunakan untuk percobaan selanjutnya yaitu untuk menentukan nilai momentum. Nilai momentum ditentukan dengan cara melakukkan eksperimen dengan memasukkan nilai dengan range 0.1 sampai dengan 1.0. Untuk nilai training *cycles* 1000 dan *learning rate* 0.3 berdasarkan percobaan sebelumnya. Berikut adalah hasil dari percobaan yang telah dilakukan untuk penentuan nilai momentum:

E-ISSN: 2548-3331 ; 145 – 154

Tabel 4 Penelitian Penentuan Nilai *Momentum*

Training Cycles	Learning Rate	Momentum	Accuracy	AUC
1000	0.3	0.1	91.09	0.969
1000	0.3	0.2	91.21	0.969
1000	0.3	0.3	90.93	0.968
1000	0.3	0.4	90.77	0.969
1000	0.3	0.5	90.82	0.968
1000	0.3	0.6	90.87	0.969
1000	0.3	0.7	90.69	0.968
1000	0.3	0.8	90.28	0.968
Training Cycles	**Learning Rate**	**Momentum**	**Accuracy**	**AUC**
1000	0.3	0.9	89.87	0.967
1000	0.3	1	52.4	0.262

Sumber: Hasil Penelitian (2016)

Berdasar eksperimen diatas, hasil klasifikasi terbaik dengan menggunakan nilai parameter training cycle=1000, learning rate=0.3 dan momentum=0.3 dengan ketepatan klasifikasi sebesar 91.21% dengan AUC 0.969. Peluang situs *phishing* dan situs non *phishing* yang mengalami klasifikasi hanya sebesar 8.79%. artinya dari 11.055 data yang salah di klasifikasikan ada sebesar 972.

Tabel 5. Tabel *confusion Matrix* NN

		C		
		Prediksi		
		Phishing	Non-*Phishing*	**Total**
Pengamatan	*Phishing*	4318	392	4710
	Non-*Phishing*	580	5765	6345
	Total	**4898**	**6157**	**11055**

Sumber: Hasil Penelitian (2016)

Dari tabel 5 menunjukkan situs *phishing* yang salah diklasifikasi sebesar 580 sebagai situs non-*phishing* dari total 4898 situs *phishing*. Sedangkan situs non-*phishing* yang salah diklasifikasikan sebesar 392 sebagai situs *phishing* dari total 6157 situs non-*phishing*. Tingkat akurasi dengan menggunakan algoritma *neural network* adalah sebesar 91,21%.

3.2 Hasil *Training* dengan *K-Nearest Neighbour*

Eksperimen selanjutnya dengan menggunakan algoritma K-Nearest Neighbor (KNN). KNN adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Teknik ini sangat sederhana dan mudah diimplementasikan. Data pembelajaran diproyeksikan ke ruang berdimensi banyak, dimana masing-masing dimensi merepresentasikan fitur dari data. Ruang ini dibagi menjadi bagian-bagian berdasarkan klasifikasi data pembelajaran. Konsep dari KNN adalah mencari nilai dari k buah data training yang jaraknya paling dekat dengan data baru yang labelnya belum diketahui (data *testing*). Yang mana untuk mendapatkan besaran jarak besaran jarak ini dihitung dengan menerapkan perhitungan jarak *Euclidean*.

K yang akan digunakan pada analisis penelitian ini yaitu 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. Berikut hasil klasifikasi dengan metode KNN:

Tabel 6 Penelitian Penentuan Nilai *k*

k	**Akurasi (%)**	**AUC**
1	90.1	0.5
2	90.33	0.903
3	90.28	0.931
4	90.29	0.947
5	90	0.954
6	90.33	0.963
7	89.21	0.964
8	90.14	0.962
9	89.59	0.961
10	89.95	0.965

Sumber: Hasil Penelitian (2016)

Dari Tabel 6 dapat dilihat bahwa nilai ketepatan klasifikasi paling optimal adalah saat k = 6, dengan ketepatan klasifikasi sebesar 90.33% dengan AUC 0.963. Artinya peluang situs *phishing* dan situs non *phishing* yang mengalami klasifikasi hanya sebesar 9.67%. artinya dari 11.055 data yang salah di klasifikasikan ada sebesar 1069.

Tabel 7 Tabel *confusion Matrix K-NN*

Tabel *confusion Matrix K-NN*

		Prediksi		
		Phishing	Non-*Phishing*	**Total**
Pengamatan	*Phishing*	4402	582	4984
	Non-*Phishing*	496	5575	6071
	Total	**4898**	**6157**	**11055**

Sumber: Hasil Penelitian (2016)

Berdasrkan tabel 7 menunjukkan situs *phishing* yang salah diklasifikasi sebesar 496 sebagai situs non-*phishing* dari total 4898 situs *phishing*. Sedangkan situs non-*phishing* yang salah diklasifikasikan sebesar 582 sebagai situs *phishing* dari total 6157 situs non-*phishing*. Tingkat akurasi dengan menggunakan algoritma *k-nn* adalah sebesar 90,33%.

4. Kesimpulan

Berdasarkan penelitian yang telah dilakukan dapat disimpulkan bahwa metode *neural network* memberikan hasil klasifikasi terbaik dengan menggunakan nilai parameter training cycle=1000, learning rate=0.3 dan momentum=0.3 dengan ketepatan klasifikasi sebesar 91.21% dengan AUC 0.969. Sedangkan metode KNN memberikan hasil performansi klasifikasi terbaik saat k = 6 dengan hasil ketepatan klasifikasi 90.33%. *Neural network* dengan *backpropagation* mampu memberikan hasil klasifikasi yang lebih baik dibandingkan dengan metode *k-nn*.

Referensi

Amato F, Lopez A, Pena-Mendez EM, Vanhara P, Hampl A. 2013. Artificial neural networks in medical diagnosis. *APPLIED BIOMEDICINE*, 47-58.

Berndtsson M, Hansson J, Lundell BO. 2008. *Thesis Projects.* London: Springer.

Cohen L, Manion L, Morrison K. 2007. *Research Methods in Education.* New York: Routledge.

Colas F, Brazdil P. 2006. Comparison of SVM and Some Older Classification Algorithms in Text Classification Tasks. *Artificial Intelligence in Theory and Practice* , 169-178.

Dawson C. 2007. *A Practical Guide To Research Methods.* United Kingdom: How To Books .

E-ISSN: 2548-3331 ; 145 – 154

Ingham J, Cadieux J, Berrada AM. 2014. E-Shopping Acceptance: a Qualitative and Meta-Analytic Review. *Information and Management*.

Jadav K, Pancha M. 2012. Optimizing Weights of Artificial Neural Networks using Genetic Algorithms. *International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARCSEE)*, 47-51.

Mohammad RM, Thabtah F, McCluskey L. 2014. Predicting *phishing* websites based on self-structuring neural network. *Neural Computing and Applications*, 443-458.

Piro P, Nock R, Nielsen F, Barlaud M. 2012. Leveraging k-NN for generic classification boosting. *Neurocomputing*, 3-9.