POLITECHNIKA GDAŃSKA
WYDZIAŁ INŻYNIERII LĄDOWEJ I ŚRODOWISKA

WILiŚ

Imię i nazwisko studenta: Grzegorz Wesołowski
Nr albumu: 127469

Studia drugiego stopnia

Forma studiów: stacjonarne

Kierunek studiów: Transport

Specjalność/profil: Systemy Transportowe

**PRACA DYPLOMOWA MAGISTERSKA**

Tytuł pracy w języku polskim:

**Opracowanie metody szacowania liczby wypadków i ofiar wypadków na odcinkach dróg z zastosowaniem sieci neuronowych.**

Tytuł pracy w języku angielskim:

**Conception of road accident prediction method using Artificial Neural Networks.**

| Potwierdzenie przyjęcia pracy | |
|---|---|
| Opiekun pracy<br><br>*podpis* | Kierownik Katedry/Zakładu<br><br>*podpis* |
| | |

Data oddania pracy do dziekanatu:

POLITECHNIKA GDAŃSKA
WYDZIAŁ INŻYNIERII LĄDOWEJ I ŚRODOWISKA

WILiŚ

## OŚWIADCZENIE

Imię i nazwisko: Grzegorz Wesołowski
Data i miejsce urodzenia: 02.10.1990, Więcbork
Nr albumu: 127469
Wydział: Wydział Inżynierii Lądowej i Środowiska
Kierunek: transport
Poziom studiów: II stopnia
Forma studiów: stacjonarne

Ja, niżej podpisany(a), wyrażam zgodę/nie wyrażam zgody* na korzystanie z mojej pracy dyplomowej zatytułowanej: *Opracowanie metody szacowania liczby wypadków i ofiar wypadków na odcinkach dróg z zastosowaniem sieci neuronowych* do celów naukowych lub dydaktycznych.[1]

Gdańsk, dnia ..... .....
*podpis studenta*

Świadomy(a) odpowiedzialności karnej z tytułu naruszenia przepisów ustawy z dnia 4 lutego 1994 r. o prawie autorskim i prawach pokrewnych (Dz. U. z 2006 r., nr 90, poz. 631) i konsekwencji dyscyplinarnych określonych w ustawie Prawo o szkolnictwie wyższym (Dz. U. z 2012 r., poz. 572 z późn. zm.),[2] a także odpowiedzialności cywilno-prawnej oświadczam, że przedkładana praca dyplomowa została opracowana przeze mnie samodzielnie.

Niniejsza(y) praca dyplomowa nie była wcześniej podstawą żadnej innej urzędowej procedury związanej z nadaniem tytułu zawodowego.

Wszystkie informacje umieszczone w ww. pracy dyplomowej, uzyskane ze źródeł pisanych i elektronicznych, zostały udokumentowane w wykazie literatury odpowiednimi odnośnikami zgodnie z art. 34 ustawy o prawie autorskim i prawach pokrewnych.

Potwierdzam zgodność niniejszej wersji pracy dyplomowej z załączoną wersją elektroniczną.

Gdańsk, dnia ..... .....
*podpis studenta*

Upoważniam Politechnikę Gdańską do umieszczenia ww. pracy dyplomowej w wersji elektronicznej w otwartym, cyfrowym repozytorium instytucjonalnym Politechniki Gdańskiej oraz poddawania jej procesom weryfikacji i ochrony przed przywłaszczaniem jej autorstwa.

Gdańsk, dnia ..... .....
*podpis studenta*

*) niepotrzebne skreślić

[1] Zarządzenie Rektora Politechniki Gdańskiej nr 34/2009 z 9 listopada 2009 r., załącznik nr 8 do instrukcji archiwalnej PG.

[2] Ustawa z dnia 27 lipca 2005 r. Prawo o szkolnictwie wyższym:
Art. 214 ustęp 4. W razie podejrzenia popełnienia przez studenta czynu podlegającego na przypisaniu sobie autorstwa istotnego fragmentu lub innych elementów cudzego utworu rektor niezwłocznie poleca przeprowadzenie postępowania wyjaśniającego.
Art. 214 ustęp 6. Jeżeli w wyniku postępowania wyjaśniającego zebrany materiał potwierdza popełnienie czynu, o którym mowa w ust. 4, rektor wstrzymuje postępowanie o nadanie tytułu zawodowego do czasu wydania orzeczenia przez komisję dyscyplinarną oraz składa zawiadomienie o popełnieniu przestępstwa.

## STRESZCZENIE

Praca dotyczy prognozowania wybranych wskaźników bezpieczeństwa ruchu drogowego oraz określenia wpływu poszczególnych cech drogi na bezpieczeństwo ruchu drogowego. Prognozowaniem objęto liczbę, gęstość i koncentrację wypadków, ofiar rannych i zabitych na podstawie określonych cech jednojezdniowych odcinków polskich dróg krajowych. Korzystając z doświadczeń zagranicznych, zastosowano sztuczne sieci neuronowe - matematyczne struktury, pozwalające modelować i prognozować szerokie spektrum zjawisk. Wykorzystano powszechnie dostępne narzędzia w postaci pakietu ANN (Artificial Neural Networks) środowiska Scilab. Zestaw cech wejściowych opisujących odcinki drogowe został określony m.in. w oparciu o analizę PCA (Principal Component Analysis), dostępną w środowisku R. Uzyskane rezultaty przeanalizowano i porównano z dotychczasowymi osiągnięciami. W pracy pokazano użyteczność sztucznych sieci neuronowych przy badaniach opisanego zagadnienia. Stworzono modele, które mogą zostać wykorzystane m.in. do prognozowania liczby wypadków na odcinkach dróg o zmienionych parametrach, w przyszłości lub na innych drogach. Innym celem pracy było określenie wpływu poszczególnych cech drogi na bezpieczeństwo ruchu drogowego, co sieci neuronowe również umożliwiają. Większość wniosków w tej kwestii pokrywa się z dotychczasowymi doświadczeniami. Stwierdzono jednak kilka interesujących faktów, takich jak: istotność rodzaju pobocza, czy niewielkie znaczenie udziału odcinków zabudowanych i zadrzewionych.

## ABSTRACT

This paper deals with prognoses of chosen road traffic safety rates and estimating the influence of different road characteristics on traffic safety. Prognoses concern number, density and concentration of accidents and casualties. They are based on real characteristics of polish single carriageway National road sections. Experience gained by worldwide researchers was used to base prognoses on Artificial Neural Networks - mathematical structures, which help to create models and make predictions of many problems. Main tool used to create the model was Scilab platform with dedicated ANN toolbox. First of all, the input sets were created using, among others, Principal Component Analysis (PCA) procedure available in R language. The results were discussed and compared to earlier works. On this basis, it can be assumed, that ANN can be useful in examined problem. Created models can be used e.g. for forecasting safety on road sections in future, road sections with changed characteristics or even on different roads. Other objective of these analysis was to define significance of different road features to road safety. It is possible thanks to ANN based models. Main conclusions where similar to earlier experiences. In other hand, some interesting facts appeared, like great importance of road shoulders and lack of significance of urban and forest sections percentage.

## Spis treści

**Wykaz ważniejszych oznaczeń i skrótów**

*PCA* – Analiza Głównych Składowych (Principal component analysis)

*MSE* – błąd średniokwadratowy (Mean Squared Error) [-]

*RMSE* – pierwiastek błędu średniokwadratowego (Root Mean Squared Error) wg wzoru 3.3. [-]

*MRE* – średni błąd względny (Mean Relative Error) wg wzoru 3.4. [%]

$r^2$ – poziom dopasowania wg wzoru 3.5. [-]

$R^2$ – współczynnik determinacji wg wzoru 3.6. [-]

*poj.* – pojazd

*os.* – osoba

*wyp.* – wypadek

*L.* – liczba

# 1. WSTĘP I CEL PRACY.

Kwestia bezpieczeństwa ruchu drogowego jest jednym z najważniejszych problemów współczesnego transportu. W Polsce najwięcej wypadków ma miejsce na odcinkach dróg krajowych. Długość sieci zamiejskich tych dróg w 2011r. wynosiła 17 460 km, co stanowiło 4,5% dróg publicznych w Polsce. Wydarzyło się na ich jednak w tymże roku niemal 8 tys. wypadków, co stanowiło 19,9% ogólnej liczby wypadków w kraju. Zginęło w nich z kolei 1513 osób, czyli aż 36,1% zabitych na wszystkich drogach [1].

Powyższe dane wskazują na konieczność zwrócenia szczególnej uwagi na zagadnienie modelowania bezpieczeństwa na odcinkach dróg krajowych w Polsce. Działania te mogą pomóc w zrozumieniu problemu i przyczynić się do znalezienia rozwiązań mogących w pozytywny sposób wpłynąć na poziom bezpieczeństwa. Dotychczas przeprowadzono liczne badania i wydano wiele opracowań w tym zakresie. Wśród najważniejszych można wymienić Narodowy Program Bezpieczeństwa Ruchu Drogowego – Gambit (*http://www.krbrd.gov.pl*) czy Atlas Bezpieczeństwa Ruchu Drogowego w Polsce – Eurorap (*http://www.eurorap.pl*). Wszelkie działania opierają się jednak w większości na tradycyjnych sposobach analizy danych i oceny ryzyka.

Tymczasem w dzisiejszych czasach, w różnego rodzaju badaniach, rosnące znaczenie mają bardziej zaawansowane metody, takie jak sieci neuronowe. Te samouczące się struktury zyskują zastosowanie w rozwiązywaniu coraz to nowych problemów. Zagraniczne badania pokazują, że również w dziedzinie bezpieczeństwa ruchu drogowego mogą się one okazać bardzo przydatne. Tym co odróżnia je od innych metod, jest możliwość uwzględnienia dużej liczby czynników i ich wzajemnego oddziaływania. Co więcej, sieci neuronowe pozwalają ze znaczną precyzją prowadzić prognozy dotyczące bezpieczeństwa transportu.

W niniejszej pracy, starano się wykorzystać doświadczenia i zaadaptować je do określonych warunków. Próba wykorzystania sieci neuronowych w modelowaniu bezpieczeństwa odcinków dróg krajowych w Polsce może okazać się odmiennym od dotychczasowych doświadczeń spojrzeniem na problem i w efekcie prowadzić do nowych wniosków. Określenie możliwości sieci neuronowych w odniesieniu do dróg krajowych, może okazać się przydatne również w kontekście innych badań, dotyczących np. odcinków dróg zamiejskich innych kategorii.

### *1.1. Cel i zakres pracy.*

Podstawowym celem niniejszej pracy jest stworzenie modelu pozwalającego możliwie najprecyzyjniej przewidywać wskaźniki bezpieczeństwa, zwłaszcza takie, jak liczba i gęstość wypadków oraz zabitych na odcinkach dróg krajowych o określonej charakterystyce. W tworzeniu modelu wykorzystano struktury sztucznych sieci neuronowych, w związku z czym kolejnym celem pracy będzie również określenie możliwości tej metody w zakresie przewidywania bezpieczeństwa na odcinkach dróg. Dodatkowo, na podstawie przeprowadzonych badań i analizy działania stworzonego modelu, możliwe będzie wyciągnięcie wniosków na temat wpływu poszczególnych czynników na bezpieczeństwo ruchu na odcinkach dróg krajowych. Ponadto, praca ma na celu określenie wskazań i wytycznych odnośnie metodyki tworzenia modelu bezpieczeństwa na drogach z wykorzystaniem sieci neuronowych.

Zakres pracy obejmuje:

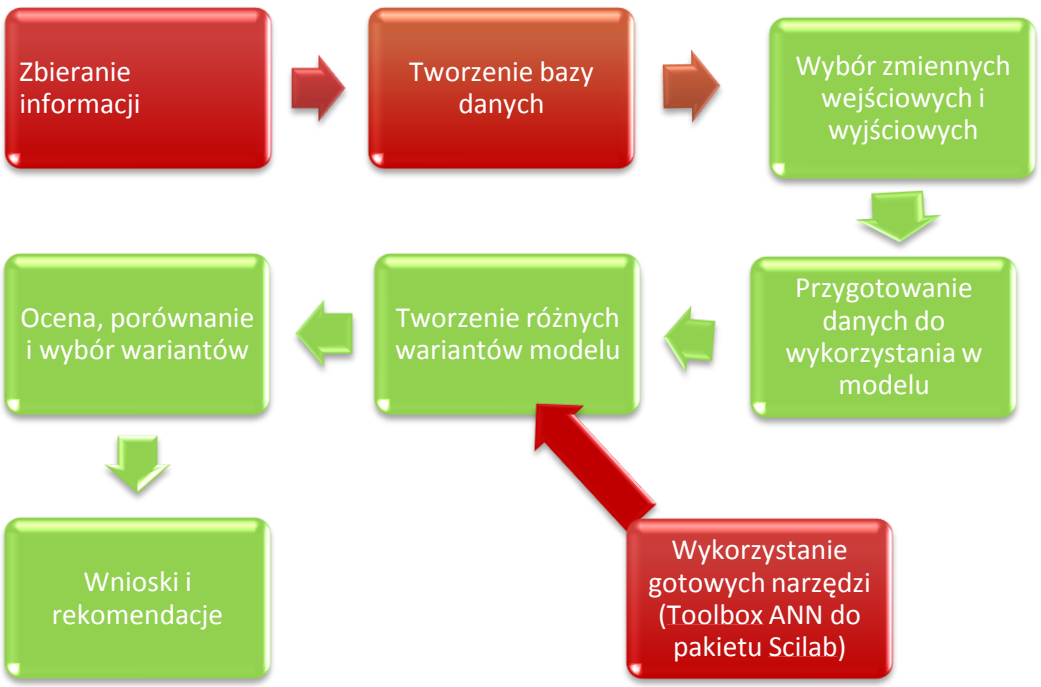
a) analizę literatury dotyczącej przede wszystkim zastosowań sztucznych sieci neuronowych w tworzeniu modeli bezpieczeństwa na odcinkach dróg oraz istniejących opracowań dotyczących bezpieczeństwa ruchu drogowego na odcinkach polskich dróg krajowych,
b) wybór zmiennych wejściowych i wyjściowych z istniejącej bazy danych wypadków drogowych na odcinkach dróg krajowych,
c) przygotowanie danych do wykorzystania w modelu m.in. poprzez podział i normalizację danych, przeprowadzenie analizy PCA,
d) stworzenie różnych wariantów modelu z wykorzystaniem sieci neuronowych. Poszczególne warianty mogą różnić się co do struktury sieci, sposobów uczeniu, doboru danych wejściowych i wyjściowych a także innych cech,
e) ocenę stworzonych modelów na podstawie określonych kryteriów, ich porównanie i wybranie najbardziej optymalnych wariantów. Wybrane rozwiązanie może również zostać porównanie z modelem nie wykorzystującym struktur sieci neuronowych,
f) analiza stworzonego modelu pod kątem określenia wpływu poszczególnych czynników na bezpieczeństwo ruchu,
g) przedstawienie wniosków obejmujących m.in. ocenę jakości i przydatności modelu, a także rekomendacje dotyczące jego zastosowań oraz najważniejsze wytyczne dotyczące tworzenia tego typu modeli.

Przy tworzeniu modelu, korzystano z gotowego oprogramowania w postaci zestawu narzędzi programistycznych - toolboxa ANN (ang. Artificial Neural Network), przeznaczonego do pakietu Scilab [23].

Praca nie obejmuje swym zakresem zbierania danych i tworzenia bazy. Dane uwzględnione w pracy pochodzą z gotowej bazy, obejmującej wypadki na zamiejskich odcinkach dróg krajowych w Polsce, o dominującym udziale odcinków jednopasmowych. Zachowano przy tym podział na odcinki przyjęty przez autorów bazy. Dane obejmują wypadki z lat 2006 – 2008.

Efektem pracy ma być przygotowanie modeli w taki sposób, by użytkownik mógł z nich korzystać, dokonując prognoz bazujących na własnych danych. Zadanie to nie obejmuje jednak tworzenie graficznego interfejsu użytkownika, dającego możliwość korzystania z modelu bez wglądu w jego strukturę.

Poszczególne fazy tworzenia modelu przedstawiono na schemacie poglądowym (rys.1.1.). Kolorem zielonym zaznaczono działania, które obejmuje niniejsza praca, natomiast na czerwono pozostałe działania, niezbędne przy tworzeniu modelu.




Rysunek 1.1. Rama logiczna pracy

# 2. SZTUCZNE SIECI NEURONOWE.

Poniższy rozdział zawiera analizę literatury związanej z tematyką pracy. Starano się przy tym poruszyć wszystkie zagadnienia, potrzebna w dalszej części opracowania.

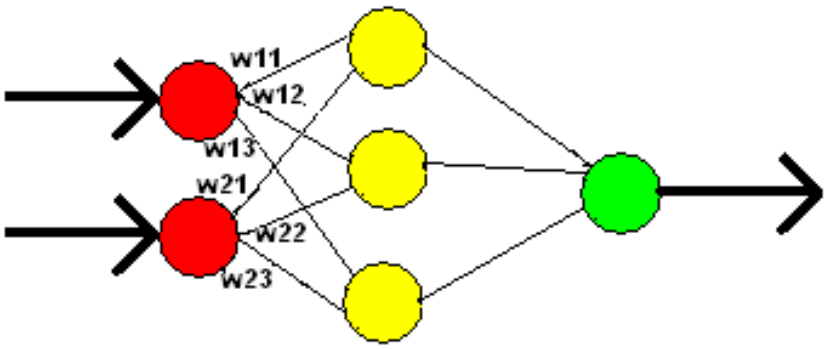
## *2.1. Definicja i zastosowania sieci neuronowych w transporcie.*

Do modelowania różnorodnych zagadnień i problemów badawczych, wykorzystuje się bardzo szeroką paletę metod i narzędzi. Wśród nich, rosnącą popularnością cieszą się sztuczne sieci neuronowe. Są to matematyczne struktury lub gotowe programy, służące przetwarzaniu sygnałów w oparciu o matematyczny model biologicznych sieci neuronowych. Struktura sztucznych sieci neuronowych przypomina więc naturalne neurony i połączenia między nimi. Połączeniom tym przypisywane są wagi, wyznaczane w procesie uczenia [2].

Sztuczna sieć neuronowa składa się z kilku warstw:

- wejściowej, przez którą dane wprowadzane są do sieci,
- ukrytej, w których odbywa się przetwarzanie danych wejściowych w wyjściowe za pomocą określonych funkcji,
- wyjściowej, do wyznaczania ostatecznej wartości wyjściowej sieci.

Na rys. 2.1. przedstawiono przykładową strukturę takiej sieci, o następujących parametrach: liczba neuronów w poszczególnych warstwach: 2 – 3 – 1. Poza tym symbole w11 – w23 – oznaczają wartości przypisywane poszczególnym połączeniom.




Rysunek 2.1. Schemat przykładowej sieci neuronowej. Opracowanie własne na podstawie [3]

Liczba neuronów w poszczególnych warstwach może być różna – w warstwie wejściowej odpowiada liczbie wprowadzanych jednocześnie wartości wejściowych. Analogicznie warstwa wyjściowa składa się z tylu neuronów, ile różnych wartości wyjściowych sieć ma określać. Z kolei część ukryta może składać się z kilku warstw o określonej liczbie neuronów. Poszczególne neurony warstwy ukrytej składają się z bloku sumowania oraz bloku aktywacji, którym jest określona funkcja [3].

Istnieją dwa zasadnicze rodzaje sieci neuronowych, różniące się sposobem uczenia sieci:

- uczenie z nauczycielem (z nadzorem)
- uczenie bez nauczyciela (bez nadzoru) – np. sieć Kohonena.

W sztucznej sieci neuronowej z nauczycielem, na wejściu podaje się oprócz pewnych cech również oczekiwane wartości wyjściowe. Podczas kolejnych faz, na etapie uczenia, sieć przypisuje wagi poszczególnym połączeniom w taki sposób, aby uzyskać wyniki jak najbardziej zbliżone do rzeczywistych. Po skończonym procesie uczenia sieć jest w stanie określać wartości wyjściowe już tylko na podstawie cech wejściowych (bez znajomości realnych wartości wyjść).

Sieć bez nadzoru nie zna oczekiwanych wartości wyjściowych. Jej zadaniem jest grupowanie danych i tworzenie logicznych klas, bądź odtworzenie na zasadzie skojarzeń całości informacji na podstawie jej znanego fragmentu [2, 3].

Wśród wad sztucznych sieci neuronowych można wymienić dość znaczny stopień skomplikowania struktur. Co więcej, użytkownikowi często ciężko jest precyzyjnie określić procesy zachodzące wewnątrz struktur sieci, a przez to zinterpretować otrzymane przez model wyniki.

Sieci neuronowej posiadają jednak wiele zalet w stosunku do innych metod matematycznych. Po pierwsze, jest to możliwość uwzględnienia w modelu wpływu kombinacji dużej liczby różnych czynników na określony wynik. Ponadto, uczenie sztucznych sieci odbywa się samoczynnie (nie wymaga programowania). Jego efektem jest m.in. zdolność do uogólniania zdobytej wiedzy. Między innymi dzięki temu, sieci neuronowe znajdują coraz szersze zastosowanie w dziedzinie transportu.

Można wyróżnić kilka podstawowych obszarów wykorzystania sieci neuronowych w tym zakresie [4, 5]:

a) związane z ruchem drogowym, np.:
- określanie przewidywanego czasu podróży na podstawie warunków ruchu,
- przewidywanie długości kolejki na skrzyżowaniach,
- krótkoterminowe prognozowanie natężenia ruchu pojazdów na skrzyżowaniach.

b) związane z infrastrukturą transportową, np.:
- wyznaczanie przepustowości dróg wielopasmowych,
- modelowanie i wykrywanie spękań nawierzchni,

c) związane z planowaniem i prognozami ruchu, np.:
- modelowanie wyboru trasy podróży,
- przewidywanie natężenia ruchu międzymiastowego,
- modelowanie rozkładu podróży na sieć transportową.

d) związane z wpływem transportu na środowisko, np.:
- przewidywanie zanieczyszczeń powietrza.

e) związane z Inteligentnymi Systemami Transportu, np.:
- wykrywanie pojazdów i ich śledzenie,
- identyfikacja pojazdów na podstawie tablic rejestracyjnych,
- wykrywanie wypadków.

f) związane z bezpieczeństwem transportu, np.:
-przewidywanie ciężkości zdarzeń drogowych z udziałem dwóch pojazdów na skrzyżowaniach z sygnalizacją świetlną,
-przewidywanie ryzyka wypadków drogowych na odcinkach międzywęzłowych dróg,
-przewidywanie liczby zdarzeń drogowych na podstawie czynników takich, jak: natężenie ruchu, warunki pogodowe, geometria drogi, ograniczenia prędkości itp.

g) pozostałe, np.:
- przewidywanie wielkości przewozów w transporcie morskim,
- przewidywanie zapotrzebowania pasażerów na transport kolejowy,
- przewidywanie czasu przyjazdu autobusów.

Wydaje się, że spośród wielu wymienionych zastosowań sieci neuronowych, do najważniejszych i posiadających największy potencjał należy zaliczyć te związane z modelowaniem bezpieczeństwa.

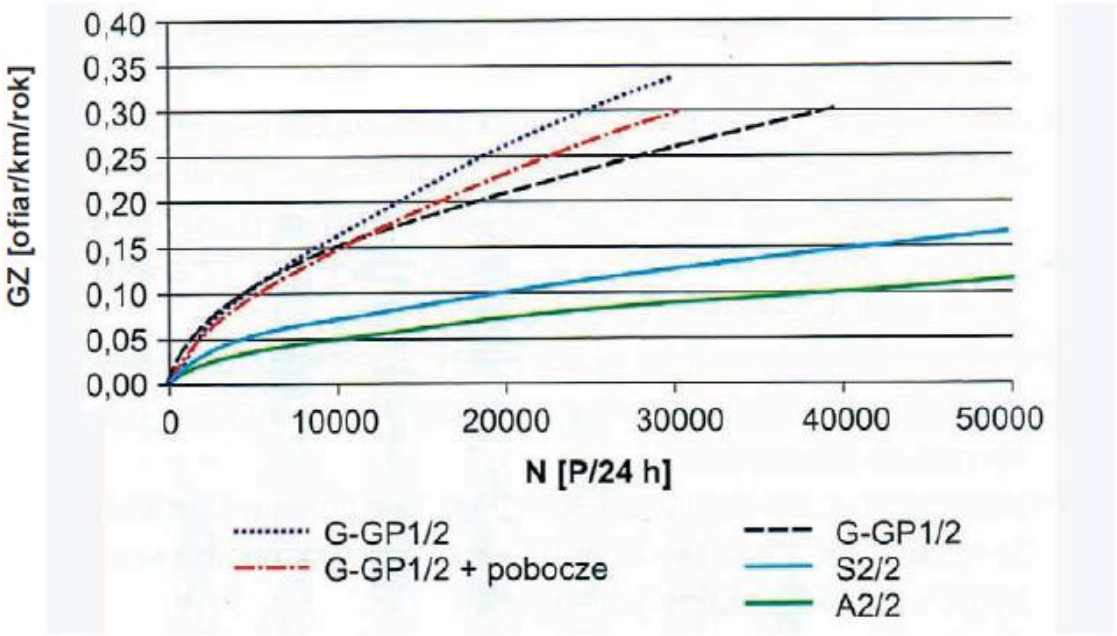
# 3. MODELOWANIE BEZPIECZEŃSTWA RUCHU DROGOWEGO.

## *3.1. Istniejące modele dotyczące bezpieczeństwa dróg krajowych w Polsce.*

Problem bezpieczeństwa na drogach krajowych w Polsce był dotąd dość szeroko poruszany [1,6,22]. Przede wszystkim, powstały programy poprawy bezpieczeństwa, takie jak „*Drogi zaufania - program ochrony życia i zdrowia ludzi na drogach krajowych*", czy Narodowy Program Bezpieczeństwa Ruchu Drogowego 2013 - 2020. Z drugiej strony, nie brakuje także prac naukowych dotyczących oceny, modelowania i prognozowania bezpieczeństwa. Niektóre z nich związane są stricte z drogami krajowymi, inne obejmują nieco szerszy zakres. W większości zbudowanych modeli nie korzystano jednak z metody sztucznych sieci neuronowych. Nieliczne wyjątki zazwyczaj dotyczą tylko wybranego aspektu bezpieczeństwa, są również ograniczone, jeśli chodzi o obszar analizy.

Jako jedną z ważnych prac dotyczących bezpieczeństwa na odcinkach dróg należy wymienić pracę W. Kustry i K. Jamroza [6]. Badacze próbowali w niej wyznaczyć najbardziej istotne cechy zwiększające ryzyko wypadków. Wykorzystano w tym celu pakiet Statistica, za pomocą którego stworzono modele oparte o metody matematyczne, takie jak regresja nieliniowa. Działano w oparciu o zestaw danych statystycznych obejmujących wypadki przypisane do miejsc na sieci dróg wraz z podstawowymi cechami odcinków dróg. W modelu najwierniej obrazującym rzeczywistość, prognoz dokonuje się na podstawie kilku czynników. Wśród czynników tych, najmocniej wpływających na gęstość zabitych na drogach, wymieniono przede wszystkim natężenie ruchu, typ drogi, udział pojazdów ciężkich i udział obszarów zabudowanych na danym odcinku drogi. Uzyskano przy tym współczynnik determinacji $R^2$ wynoszący 0,47 dla liczby zabitych na drogach jednojezdniowych i 0,44 dla gęstości. Jako inny istotny czynnik wpływający na powstawanie wypadków, podaje się prędkość ruchu i liczbę skrzyżowań [6].

Problem bezpieczeństwa na drogach krajowych pod kątem przyczyn wypadków wspomniany również został w pracy M. Budzyńskiego i W. Kustry [22]. Przedstawiono w niej zależność (rys. 3.1.), z której wynika, że natężenie ruchu i rodzaj drogi w sposób wyraźny wpływają na bezpieczeństwo ruchu, podobnie jak procentowy udział pojazdów ciężkich [17]. Wnioski tego typu pojawiają się w większości prac badawczych.

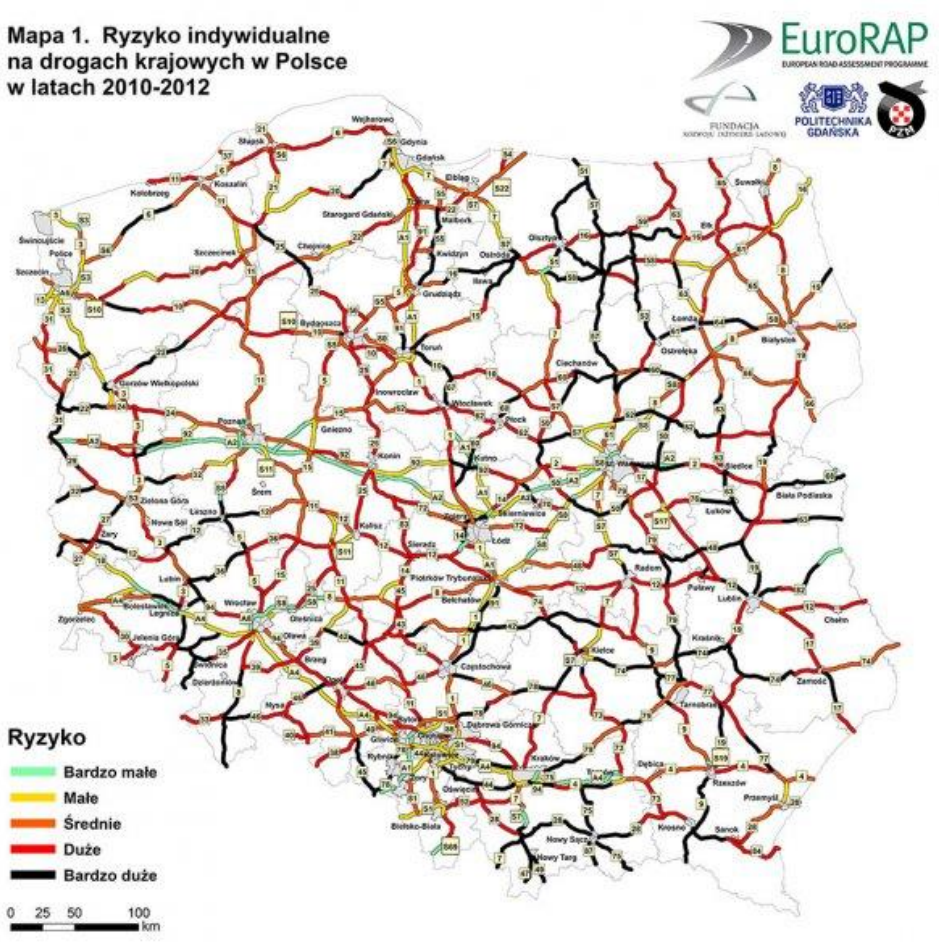



Rysunek 3.1. Wpływ natężenia ruchu i typu drogi na gęstość wypadków na drogach krajowych w terenie niezabudowanym przy średnim udziale pojazdów ciężarowych. Źródło: [17]

Modele dotyczące analizy i przewidywania buduje się zazwyczaj w oparciu o wielokrotną regresję liniową lub nieliniową. Ponadto wykorzystuje się szeregi czasowe, czy diagramy drzewiaste.

Jedna z niewielu prac odnoszących się do polskich dróg krajowych, w której model oparto o sztuczne sieci neuronowe jest autorstwa M. Nowakowskiej („*Analiza typologiczna wypadków drogowych z wykorzystaniem sztucznej sieci neuronowej Kohonena*"). Autorka skupiła się w niej na jednojezdniowych, dwukierunkowych odcinkach dróg zamiejskich, wyłącznie na terenie województwa Świętokrzyskiego. Kolejnym ograniczeniem był rodzaj zdarzenia: brano pod uwagę jedynie wypadki bez pieszych, z udziałem tylko jednego pojazdu silnikowego (np. uderzenie w drzewo). Analiza dotyczyła lat 2004 – 2007. W trakcie prac stworzono model szczegółowy (niezagregowany). Oznacza to, że każde zdarzenie analizowane było oddzielnie, jako pojedynczy rekord. W związku z tym, zastosowano model sieci neuronowej bez nauczyciela – tzw. sieć Kohonena. Jej zadaniem było pogrupowanie wypadków na kilka klas o podobnych cechach. Opisanie podobieństw wewnątrz klas może pomóc w zrozumieniu istoty i przyczyny poszczególnych typów grup wypadków. Jest to podejście znacząco różne od założeń przyjętych w niniejszym opracowaniu [6].

Na koniec warto wspomnieć o Europejskim Programie Oceny Ryzyka na Drogach (EuroRAP). W ramach tego programu, opracowuje się mapy ryzyka dla dróg krajowych w Polsce. Działania polegają na zestawieniu statystycznym wypadków na poszczególnych odcinkach dróg i zakwalifikowaniu tych odcinków do grup ryzyka. Analizuje się przy tym liczbę ofiar ciężko rannych i zabitych w stosunku do pracy przewozowej na odcinku (ryzyko indywidualne) oraz w stosunku do długości odcinka (ryzyko społeczne). Powstałe na tej podstawie mapy mogą służyć zarówno zarządcom dróg, jak i przewoźnikom oraz osobom prywatnym, poruszającym się po sieci dróg krajowych w Polsce. Zasadność opracowań EuroRAP jest bezdyskusyjna, należy jednak zauważyć, że wyniki w niewielkim stopniu odnoszą się do przyczyn wypadków, skupiając się wyłącznie na ich skutkach. Nie są więc uwzględniane chociażby cechy poszczególnych odcinków dróg. W związku z tym, zestawienie efektów programu EuroRAP z przewidywanymi rezultatami niniejszej pracy może prowadzić do nowych wniosków.




Rysunek 3.2. Ryzyko indywidualne na drogach krajowych w Polsce w latach 2010 – 2012. Źródło: eurorap.pl

## *3.2. Istniejące modele dotyczące bezpieczeństwa na odcinkach dróg, wykorzystujące sieci neuronowe.*

Wśród wielu zastosowań sieci neuronowych w transporcie, nie brakuje tych, które w sposób bezpośredni dotyczącą modelowania bezpieczeństwa transportu. Dominują przy tym badania zagadnień dotyczących transportu drogowego [4,5]. Sporym zainteresowaniem badaczy na przestrzeni ostatnich lat cieszyły się zwłaszcza skrzyżowania dróg oraz odcinki dróg (pomiędzy węzłami). Modele obu typów, jakkolwiek pokrewne, różnią się nieco, zwłaszcza co do doboru czynników wejściowych. Ponadto wypadki na odcinkach dróg i na skrzyżowaniach posiadają znacząco różne przyczyny i charakterystykę [7]. Dlatego też, w niniejszej pracy ograniczono się do opracowań dotyczących odcinków dróg.

W kolejnych punktach dokonano analizy poszczególnych aspektów wybranych, w większości zagranicznych, badań. Analiza ta może okazać się pomocna w budowie modelu dla polskich dróg krajowych.

### *3.2.1. Dobór czynników wyjściowych.*

Czynniki wyjściowe, czyli wyjścia sieci neuronowej, albo zmienne zależne to elementy, których wartość ma być przewidywana na podstawie dotychczasowych obserwacji. Do uczenia sieci niezbędna jest znajomość wartości tych cech dla określonej liczby rekordów danych wejściowych. Dopiero po przeprowadzeniu procesu uczenia sieci neuronowej, możliwe jest szacowanie wartości wyjściowych w przyszłości, bądź dla zmienionych warunków na wejściach.

Pierwszym zagadnieniem, które powinno być rozstrzygnięte jest wybór między modelem zagregowanym, a modelem szczegółowym. Model zagregowany, to taki w którym czynniki wyjściowe mają charakter ilościowy. Zmienne wyjściowe tego modelu, podobnie jak zmienne wejściowe, są wynikiem kategoryzacji lub agregacji, czyli grupowania [9].

Przykładowo, w modelu zagregowanym, wiele wypadków, które miały miejsce na jednym odcinku będzie traktowane, jako jeden rekord z danymi dotyczącymi cech tego odcinka. Zmienną celu będzie w tym wypadku liczba wypadków, ofiar rannych, czy ofiar zabitych w określonym czasie.

Jako miary stosuje się przy tym gęstość wypadków, rannych lub zabitych wyrażoną w wypadkach/km lub rannych/km [8]. Można również wyniki przedstawiać, jako wartości bezwzględne. W przypadku modelu dla dróg na Sycylii, ograniczono się do prostego wskaźnika – łącznej liczby wypadków [9].

Drugi rodzaj modeli, to modele szczegółowe. W tym przypadku, każde zdarzenie analizowane jest z osobna, zmienną celu jest cecha pojedynczego zdarzenia, najczęściej jego ciężkość. Przy podejściu tym przewiduje się skutki zdarzenia, jakie spowodowane będą przez określoną kombinację czynników. Czynnik wyjściowy jest więc opisany w sposób jakościowy, a nie ilościowy. Zwykle dokonuje się podziału na: wypadki, czyli zdarzenia, które skutkowały wystąpieniem ofiar rannych lub zabitych, a także kolizje, których efektem były jedynie straty materialne. Można zastosować również wskaźnik, którym będzie odsetek zdarzeń, w których przy danym zestawie warunków wejściowych wystąpią ofiary ranne lub zabite [10].

Możliwy jest także bardziej szczegółowy podział. Aby móc precyzyjniej wydzielić grupy skutków, trzeba dysponować dokładną bazą danych, w której znajduje się odpowiednio duża liczba zdarzeń z poszczególnych grup. Badacze z Oklahoma State University podzielili skutki zdarzeń na aż 5 grup (brak rannych, 3 stopnie ciężkości ran, zabici). Rezultaty tych badań nie były jednak w pełni satysfakcjonujące [11].

Jeśli chodzi o liczbę cech wyjściowych, zarówno w przypadku modeli zagregowanych, jak i niezagregowanych wynosiła ona zwykle 1 lub 2. Nie ma jednak przeszkód, aby dokonywać jednoczesnej prognozy większej liczby zmiennych wyjściowych [10].

Na podstawie dotychczasowych doświadczeń można stwierdzić, że modele bezpieczeństwa odcinków dróg tworzy się z myślą o tym, aby określić najbardziej niebezpieczne miejsca, albo żeby przewidywać przyszły stan bezpieczeństwa. Poza tym, na podstawie modelu można próbować określić najbardziej istotne czynniki i warunki, które wpływają na występowanie zdarzeń drogowych lub zwiększenie ich ciężkości. Dysponując taką wiedzą, można podjąć konkretne działania prewencyjne związane już z infrastrukturą lub organizacją ruchu.

*3.2.2. Klasyfikacja i dobór czynników wejściowych wybranych modeli.*

Jedną z podstawowych kwestii, jakie należy rozstrzygnąć przy budowie sieci neuronowej jest dobór zmiennych wejściowych (objaśniających) modelu. Zależy on przede wszystkim od tego, jakimi danymi dysponuje badacz oraz jakie efekty chce uzyskać. Właściwe zestawienie warstwy wejściowej sieci neuronowej jest kluczowe z punktu widzenia otrzymania dokładnych i wiarygodnych wyników. Pomaga też w lepszym zrozumieniu badanego problemu oraz efektywniejszym i szybszym uzyskiwaniu rezultatów [12].

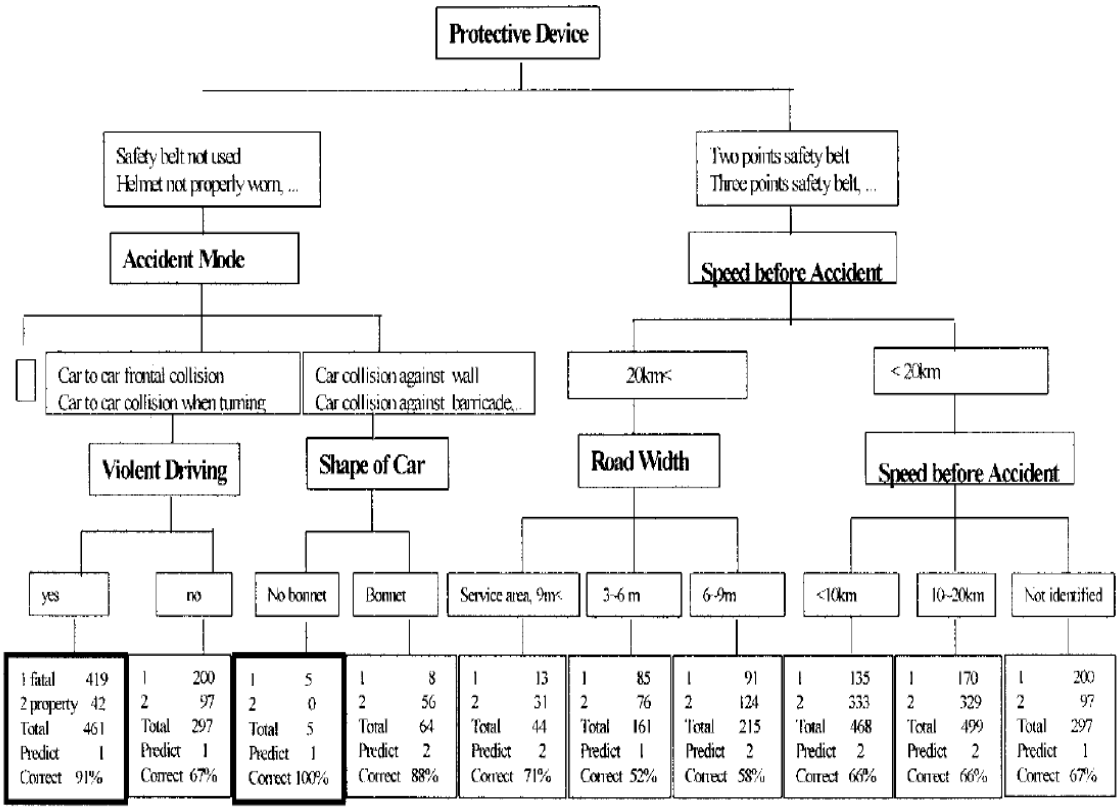
W istniejących opracowaniach na temat modeli bezpieczeństwa wykorzystujących sieci neuronowe, rzadko prezentowano zastosowaną metodykę selekcji czynników wejściowych. Dobór uwzględnianych czynników opierał się najczęściej na doświadczeniu własnym badaczy, oraz wnioskach wyciągniętych z podobnych projektów przeprowadzonych w przeszłości.

W niektórych przypadkach zastosowano jednak bardziej formalne metody. Południowokoreańscy badacze skorzystali z testów zgodności $\chi 2$ aby wstępnie wyselekcjonować 22 spośród ponad 70 dostępnych czynników. Wadą tego testu jest brak uwzględniania współzależności różnych elementów. W dalszej kolejności, za pomocą metody drzewa klasyfikacyjnego z 22 potencjalnych wejść, ograniczono się do 5, które mogły mieć najbardziej istotny wpływ na ciężkość wypadków. Ostatecznie zastosowano je w warstwie wejściowej modelu [13]. Fragment drzewa klasyfikacyjnego stworzonego w tym modelu, przedstawiono na rys. 3.3.

Badania wykazują, że drzewa klasyfikacyjne są jednym z najbardziej efektywnych sposobów klasyfikacji i doboru zmiennych. Każda zmienna reprezentowana jest przez węzeł, a gałęzie to wartości tej zmiennej. Na końcu drzewa przypisuje się przewidywaną wartość dla danej ścieżki doboru zmiennych. Na podstawie rozkładu tych wartości określa się, które ze zmiennych są najbardziej istotne [12].

W innym projekcie do selekcji czynników wejściowych zastosowano „best subset algorithm” (algorytm najlepszych podzbiorów). Na podstawie 3 kryteriów: Kryterium informacyjnego Akaikego (AIC), Mallows’ Cp, i Kryterium informacyjnego Bayesa (BIC) ograniczono liczbę wejść z 7 do 4 (szerokość pasów drogowych i pasów awaryjnych, gęstość wjazdów, lokalizacja drogi). Wadą tego algorytmu jest ograniczona liczba zmiennych, które można poddać analizie.

W przypadku konieczności przebadania istotności większej liczby czynników, autorzy proponują procedurę regresji krokowej [14]. Do eliminacji zbędnych lub redundantnych zmiennych można wykorzystać również algorytmy CFS (Correlation based Feature Selection), FCBF (Fast Correlation Based Filter algorithm), czy MIFS (Mutual Information Feature Selector) [12].



Rysunek 3.3. Fragment drzewa decyzyjnego wykorzystanego do selekcji czynników wejściowych modelu przez podział zdarzeń na wypadki (1) i kolizje (2). Źródło: [13]

Analizując dotychczasowe badania z zakresu modelowania bezpieczeństwa odcinków dróg, można stwierdzić dużą różnorodność uwzględnianych czynników. Zazwyczaj jednak należą one do jednej z następujących grup [9,10]:

- określające cechy ruchu – np. natężenie ruchu, prędkość
- określające warunki otoczenia – np. geometria drogi, stan nawierzchni, pogoda, oświetlenie, miejsce i czas wypadku
- dotyczące wypadków – np. liczba wypadków, liczba ofiar rannych i zabitych, rodzaj wypadku, rodzaj pojazdu
- określające sprawcę wypadku – np. płeć, wiek, wpływ alkoholu.

Do właściwego przeprowadzenia procesu uczenia się sieci, potrzebna jest duża ilość jednolitych i kompletnych informacji na temat zdarzeń. Liczba ta waha się od kilkuset do nawet kilkunastu tysięcy rekordów (z drugiej strony, zbyt duża liczba danych może doprowadzić do niekorzystnego, tzw. uczenia się na pamięć [15]). W związku z tym, przy planowaniu badań, należy wziąć pod uwagę fakt, że nie zawsze możliwe jest uzyskanie wystarczająco dużej i pewnej bazy danych odnośnie danego czynnika. Przykładowo, w pracy nad modelem pod wodzą M. Miao, stwierdzono znaczny wpływ prędkości pojazdów na powstawanie wypadków. Mimo świadomości faktu, naukowcy nie mogli wziąć pod uwagę tej zmiennej, ze względu na niekompletność danych – ponad 2/3 zdarzeń w bazie danych nie miała określonej prędkości pojazdów w chwili zdarzenia [11].

Istotną kwestią jest nie tylko rodzaj zmiennych wejściowych, ale także ich liczba. Co ważne, zwiększanie tej liczby wcale nie musi prowadzić do poprawienia działania sieci. Wręcz przeciwnie, zbyt dużo zmiennych wejściowych może utrudniać proces uczenia. Niestety, nie istnieje sposób na dokładne wyznaczenie optymalnej liczby wejść [15].

Biorąc powyższe stwierdzenia pod uwagę, nie jest zaskakujący fakt, iż poszczególne modele praktyczne różnią się dość znacząco co do liczby wejść i doboru zmiennych. W badaniach przeprowadzonych w chińskim Harbinie wzięto pod uwagę tylko cztery wartości wejściowe: natężenie ruchu, prędkość dopuszczalna na danym odcinku, szerokość pasa ruchu, wykorzystanie przepustowości. Dane te wystarczyły, aby ze znaczną dokładnością przewidywać gęstość wypadków. Ocena przeprowadzona po testowaniu modelu wykazała, że wśród nich najbardziej znaczący był czynnik natężenia ruchu, a najmniej istotna szerokość pasa [8].

W przypadku modelu stworzonego przez badaczy irańskich uwzględnia się aż 25 zmiennych. Wśród nich również są prędkość, natężenie ruchu i szerokość pasa, a oprócz tego: płeć i wiek kierowcy, rodzaj i przyczyna zdarzenia, typ pojazdu, warunki świetlne, miejsce zdarzenia. Liczba 25 wynika ze sposobu opisywania tych czynników (szerzej to zagadnienie opisano w podpunkcie 2.3.3) [10]. W modelu przewidującym liczbę wypadków na odcinkach sycylijskich autostrad, warstwa wejściowa składa się z 7 neuronów. Stanowiły ją współczynniki uwzględniające stan nawierzchni, warunki pogodowe, warunki ruchu (natężenie), a także występowanie na danym odcinku: łuków pionowych i poziomych, wiaduktów i tuneli [9].

Dodatkowe czynniki brane są pod uwagę w modelach badających ciężkość wypadków. We wspomnianym wcześniej koreańskim modelu, dodatkowymi czynnikami, które uwzględniono (poza prędkością, czy szerokością drogi) były również: rodzaj wypadku, kształt pojazdu oraz zastosowane urządzenia bezpieczeństwa biernego, jak pasy bezpieczeństwa, kaski. Czynniki te nie mają wpływu na powstanie zdarzenia, ale są kluczowe przy badaniu skutków wypadku [13]. Zestawienie użytych najważniejszych czynników wejściowych i wyjściowych dla wybranych badań, przedstawiono w tabeli.

Tabela 3.1. Zestawienie zmiennych w wybranych modelach dotyczących bezpieczeństwa na odcinkach dróg. Źródło: opracowanie własne.

| Autor: | Rezaie | Bosurgi [9] | Miao [11] | Zheng | Sohn | Hosseinpour [14] | Chang |
|---|---|---|---|---|---|---|---|
| **Zmienna wejściowa** | | | | | | | |
| natężenie ruchu | x | | | x | | | x |
| prędkość | x | | | x | x | | |
| stan nawierzchni | | x | x | | | | |
| geometria drogi | x | x | | x | x | x | x |
| inne cechy drogi | x | x | | | | x | x |
| cechy sprawcy | x | | x | | | | |
| cechy pojazdów | x | | x | | x | | x |
| warunki pogodowe | x | x | x | | | | x |
| rodzaj wypadku | x | | | | x | | |
| **Zmienna wyjściowa (funkcja celu)** | | | | | | | |
| liczba wypadków | | x | | | | | |
| gęstość wypadków | | | | x | | | x |
| ciężkość wypadków | | | x | | x | | |
| odsetek zdarzeń z ofiarami | x | | | | | x | |

Podsumowując, można stwierdzić, że klasyfikacja i dobór czynników wejściowych jest ważnym etapem procesu tworzeniu modelu, jednak często opiera się ona na wiedzy badaczy i nie jest poprzedzona dodatkowymi badaniami. W niektórych przypadkach, posłużono się określonymi metodami matematycznymi w celu wyselekcjonowania najbardziej istotnych czynników. Pozwala to dokładniej przewidywać wartości oczekiwane, łatwiej budować model i lepiej go rozumieć. Najczęściej uwzględnianymi czynnikami, wydają się być: prędkość, natężenie ruchu, szerokość drogi, a w przypadku badań ciężkości zdarzeń również typ wypadku i rodzaj stosowanych zabezpieczeń. Niestety, nie zawsze możliwe jest zebranie danych, które obejmowałyby wszystkie wymienione zagadnienia. Aby uczynić model dokładniejszy, można stosować różnorakie dodatkowe wskaźniki – w zależności od ich dostępności.

*3.2.3. Sposoby pozyskiwania i obróbki danych statystycznych.*

Do zbudowania modelu bezpieczeństwa drogowego opartego na sieciach neuronowych niezbędne jest posiadanie odpowiednio dużego zbioru danych. Najczęściej wykorzystuje się dane historyczne zebrane w ciągu kilku poprzednich lat. Korzysta się przy tym z dostępnych już baz, gdyż samodzielne zbieranie danych przez badaczy byłoby zbyt pracochłonne i długotrwałe. Podstawowym źródłem danych są policyjne statystyki. Zawierają one zwykle informacje o uczestnikach wypadku, jego skutkach i przebiegu. Aby poszerzyć wiedzę o warunkach panujących podczas poszczególnych zdarzeń drogowych, można odnieść się również do innych źródeł. Przykładowo, chińscy naukowcy skorzystali z przeprowadzanych niezależnie pomiarów ruchu drogowego, aby poznać natężenie na poszczególnych odcinkach. Dane o cechach odcinków (długość, rodzaj nawierzchni, dopuszczalna prędkość itp.) pozyskuje się najczęściej z odpowiednich urzędów [8].

Do dalszych badań wykorzystuje się zbiór danych z kilku lat – np. 2, 5, 6. Analizowane badania dotyczące odcinków dróg różniły się co zasięgu. Niektórzy badacze zajmowali się obszarem całego kraju (Stany Zjednoczone [11]), podczas gdy inni ograniczali się do regionu (Sycylia [9]) lub nawet jednego miasta (Harbin [8]). Zebrane dany dotyczyły odcinków o równej długości – 1km [9,14], bądź nieregularnych [8].

Pozyskane bazy danych nie nadają się zwykle do bezpośredniego wykorzystania w sieciach neuronowych. Przede wszystkim należy zmienne jakościowe zamienić na ilościowe. Badacze irańscy zastosowali w tym celu osobne wejścia sieci neuronowej. Przykładowo, za określenie przyczyny wypadku odpowiada w tym modelu aż 8 neuronów wejściowych. W każdym przypadku tylko jeden z nich posiada wartość 1, podczas gdy pozostałym przypisuje się zera. Jeśli zdarzenie spowodowane było niezachowaniem bezpiecznego odstępu, będzie to wejście C1, jeśli utratą panowania nad pojazdem neuron C6 itd. Podobnego zabiegu dokonano dla wyjść:

- wyjście Z1 – wypadek (ofiary ranne lub zabite) – wartość 1 (prawda) lub 0 (fałsz)

- wyjście Z2 – kolizja (tylko straty materialne) – wartość 1 (prawda) lub 0 (fałsz).

Aby podejście tego typu było możliwe, należy oczywiście odpowiednio zmodyfikować istniejącą bazę danych [10].

Dostępne dane należy podzielić w sposób losowy na dwa zbiory: tzw. uczący i testowy. Zbiór uczący jest większy i zawiera ok. 60 - 75% rekordów. Pozostałe służą, jako zbiór testowy [8,16].

Przygotowanie danych do modelu może polegać na ograniczeniu ich liczebności. Podczas budowy modelu dotyczącego amerykańskich dróg, stwierdzono, że największy odsetek zabitych posiadają zderzenia pojazdów jadących z naprzeciwka. W związku z tym, ograniczono się tylko do badania tej grupy zdarzeń. Następnie wyodrębniono z niej zderzenia czołowe, odrzucając pozostałe, które stanowiły łącznie zaledwie 1,3% zdarzeń. W ten sposób „ręcznie" stworzono klasę, którą poddano analizie z wykorzystaniem sieci neuronowych [11].

Bardziej zaawansowane metody eksploracji danych (ang. data mining), obejmują analizę skupień (ang. data clustering). Algorytm ten, polega na grupowaniu elementów na klasy, zawierające podobne do siebie elementy. Przykładem algorytmu z zakresu analizy skupień, są drzewa decyzyjne. W podpunkcie 2.3.2. wspomniano o ich zastosowaniu przy doborze czynników wejściowych. Ta i inne metody grupowania można wykorzystać także na kolejnym etapie budowy modelu, już po wstępnej selekcji zmiennych. Dzięki nim, można lepiej zrozumieć zależności między zmiennymi oraz zredukować ich liczbę przez łączenie ich w kategorie [17].

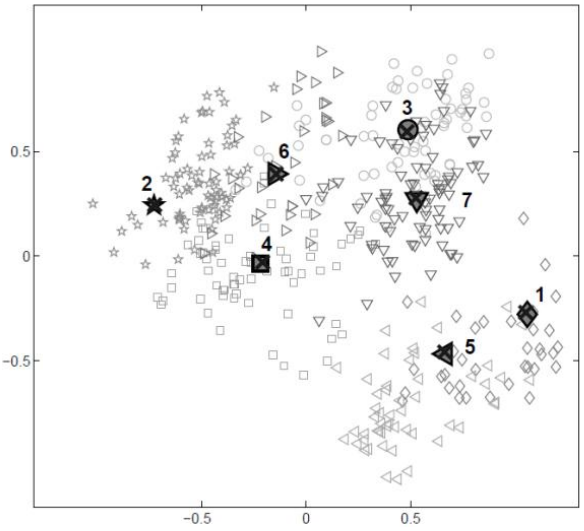
Inną metodą grupowania danych jest metoda K-średnich. W przypadku wypadków drogowych zastosowali ją naukowcy fińscy. Polega ona na wstępnym podziale danych na kilka „populacji" (w tym przypadku 7), a następnie poprawianiu podziału w kolejnych krokach. Poprawianie opiera się na przenoszeniu elementów do innych klas w taki sposób, by wariancja wewnątrz każdej klasy była minimalna. We wspomnianym przykładzie, 10-krotnie dokonywano wstępnego rozmieszczenia klas, a następnie po 100 iteracjach wybrano najlepszy podział. Przykładowo:

Klasa 1 – wypadki na drogach wielojezdniowych o dużej prędkości dopuszczalnej i natężeniu ruchu, w większości powstałe w godzinach szczytu, o przeciętnej ciężkości skutków

Klasa 2 – wypadki w dużym stopniu z udziałem nietrzeźwych kierowców, mające miejsce na drogach o mniejszym znaczeniu, w weekendy, wieczorem

Klasa 3 – zderzenia ze zwierzętami; grupa o najmniejszym ryzyku odniesienia obrażeń, większy niż w innych grupach odsetek młodych kierowców itd. [17].

Wyniki grupowania zaprezentowano graficznie, korzystając z techniki PCA (Analiza Głównych Składowych). Zgodnie z nią, wszystkie rekordy umieszcza się w przestrzeni k-wymiarowej, gdzie k to liczba zmiennych. Otrzymany układ przekształca się w taki sposób, aby maksymalizować wariancję kolejnych współrzędnych. Wyniki zaprezentowano na rys. 3.4. Do wykonania wykresu wykorzystano 0,5% przypadków. Kolejne cyfry oznaczają poszczególne klasy.



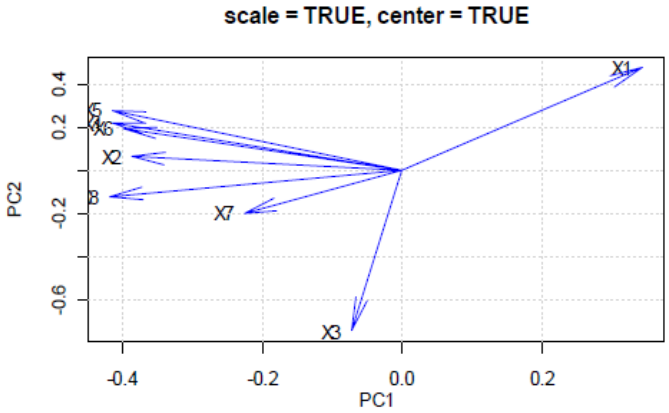

Rysunek 3.4. Wykres danych pogrupowanych na 7 klas zdarzeń, z wykorzystaniem metody PCA. Źródło: [17]

Sama metoda PCA (ang. Principal Component Analysis) również może posłużyć do redukcji wymiarowości zbioru [18]. W jej wyniku uzyskuje się nowe zmienne, zwane składowymi głównymi. Stanowią one kombinację liniową rzeczywistych cech, jednak same nie posiadają interpretacji fizycznej. Do dalszych analiz bierze się pod uwagę określoną, ograniczoną liczbę składowych głównych. Przy tym wybiera się te spośród składowych, które związane są z największą możliwą wariancją. Pozwala to w najdokładniejszy sposób odzwierciedlić zmienność występującą w oryginalnym zbiorze danych, ograniczając jednocześnie liczbę zmiennych.

Dla zilustrowania efektów zastosowania metody PCA, można wykonać dwa podstawowe rodzaje wykresów: wykres wektorów zmiennych i wykres wariancji objaśnianych przez kolejne składowe główne.

Wykres zmiennych za pomocą wektorów przypisanych do poszczególnych, oryginalnych zmiennych pokazuje ich wpływ na dwie najważniejsze składowe główne (PC1 i PC2) w przestrzeni dwuwymiarowej. Przykład wykresu zmiennych przedstawiono na rys. 3.5.

Z wykresu tego typu można wyciągnąć szereg wniosków:
- kierunek wektora i jego długość obrazuje wpływ zmiennej na składową główną
- bliskie położenie wektorów zmiennych wskazuje na ich silną dodatnią korelację, a przeciwne skierowanie korelację ujemną.
- prostopadłość wektorów sugeruje brak korelacji.




Rysunek 3.5. Przykładowy wykres położenia wektorów zmiennych względem 2 głównych składowych. Źródło: [18]

Drugi typ wykresu pokazuje wpływ kolejnych składowych głównych na wariancję. Jego analiza pomaga dobrać minimalną liczbę składowych, przy której nastąpi niewielka utrata informacji.

Podsumowując, sposoby pozyskiwania danych były podobne w przypadku wszystkich badań. W większości korzystano z gotowych baz danych. Znaczące różnice można natomiast zaobserwować, jeśli chodzi o poziom przetworzenia tych danych. W niektórych przypadkach dokonywano tylko podstawowego przygotowania – normalizacji, usunięcia zbędnych i niepełnych danych. Kiedy indziej, zastosowano zaawansowane metody eksploracji danych, których efekty niekiedy mogły posłużyć wyciąganiu daleko idących wniosków, praktycznie bez konieczności korzystania z sieci neuronowych. Metodą, która może przynieść szczególnie interesujące efekty, jest grupowanie danych z wykorzystaniem algorytmu k-średnich i Analizy Głównych Składowych (PCA).

*3.3.4. Struktura wybranych modeli.*

Każdy z wymienionych wcześniej modeli posiada ściśle określoną strukturę sieci neuronowej, przy czym występują między nimi pewne różnice. Podstawową cechą każdej sieci jest liczba neuronów w poszczególnych warstwach. W warstwie wejściowej jest ona równa liczbie zmiennych, które zdecydowano się wziąć pod uwagę. Warstwa wyjściowa składa się najczęściej z jednego lub dwóch neuronów odpowiadających zmiennym celu (wyjściowym). Pomiędzy nimi znajdują się warstwy ukryte, zwykle jedna lub dwie. Liczba neuronów w tych warstwach także może się różnić. Zbyt duża lub zbyt mała liczba neuronów w warstwie ukrytej może prowadzić do osiągania gorszych rezultatów [15].

Istnieją wzory matematyczne, za pomocą których można szacować liczbę neuronów w warstwach ukrytych. Uzależniają one tą wielkość od liczby neuronów wejściowych i wyjściowych [15]:

$$n_{ukr} = \frac{n_{wej}}{2} + n_{wyj} \qquad (3.1.)$$

lub:

$$n_{ukr} = \sqrt{n_{wyj} * n_{wej}} \qquad (3.2.)$$

Jeśli chodzi o rozwiązania zastosowane w praktyce, Irańscy badacze przygotowali do modelowania jednego zagadnienia aż 18 sieci o różnej strukturze, przy czym warstwa ukryta zawiera od 4 do 23 neuronów [10]. Warto jednak zauważyć, że w tym przypadku pierwsza warstwa składała się z aż 25 neuronów. W przypadku mniejszej liczby zmiennych, stosuje się raczej mniej neuronów również w warstwie ukrytej. I tak w badaniach Sohna – 6 neuronów wejściowych, 3 i 2 neurony w dwóch warstwach ukrytych i jeden neuron na wyjściu [13], a u Borsugiego: 7-5-1 [9].

Modele, których zadaniem jest prognozowanie liczby zdarzeń, korzystają z wielowarstwowych sieci nieliniowych. Algorytmem stosowanym w praktycznie wszystkich przypadkach jest propagacja wsteczna. Z jej wyborem wiąże się rozstrzygnięcie, co do wyboru funkcji aktywacji. Funkcja aktywacji, to funkcja za pomocą której obliczane są wartości wyjścia sieci. W opisywanych przykładach, deklarowano zazwyczaj użycie tangensa hiperbolicznego (tanh) [9,10].

*2.3.5. Sposoby oceny jakości modeli.*

Ostatnim etapem pracy nad modelem jest jego ocena. Elementu tego nie pominięto również w modelach dotyczących bezpieczeństwa na odcinkach dróg i wykorzystujących sieci neuronowe. Ocena w tych przypadkach dotyczy zarówno jakości całego modelu, jak i wpływu poszczególnych elementów na osiągane wyniki.

Podstawą do oceny dokładności prognoz jest najczęściej błąd średniokwadratowy (MSE) lub pierwiastek z błędu średniokwadratowego (RMSE). Obrazuje on różnicę między wartością estymowaną, a rzeczywistą. Wartość błędu średniokwadratowego można obliczyć z następującego wzoru:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \qquad (3.3.)$$

gdzie: *N* oznacza liczbę obserwacji, $y_i$ – wartość rzeczywistą funkcji celu dla obserwacji i, a $\hat{y}_i$ - wartość przewidywaną dla tego punktu [8].

Monitorowanie wartości tego błędu jest konieczne już w fazie uczenia sieci. Kiedy RMSE osiągnie oczekiwaną wartość, można przerwać proces uczenia. Nadmierne wydłużanie uczenia sieci, może pogorszyć osiągane wyniki. Inaczej mówiąc, liczba iteracji w procesie uczenia powinna zależeć od osiąganych wartości RMSE. Sprawdzanie błędu średniokwadratowego należy kontynuować także w fazie testowania [4].

Aby poprawić czytelność rezultatów, można policzyć również średni błąd względny (MRE), wyrażony w procentach [14]:

$$MRE\ [\%] = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{\hat{y}_i} * 100\%\right| \qquad (3.4.)$$

Inny wskaźnik zastosowany przez Zhenga [8] to poziom dopasowania ($r^2$). Jego wartość opisuje wzór:

$$r^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(\hat{y}_i)^2} \qquad (3.5.)$$

Przykład dobrze zaprojektowanego modelu pokazuje, że można osiągnąć wartości $r^2$ na poziomie 0,988, a MRE ok. 20% [8].

Do oceny różnych modeli, szczególnie wykorzystujących regresję liniową jest natomiast współczynnik determinacji $R^2$. Określa on stopień zależności pomiędzy zmienną objaśnianą, a objaśniającą (uzyskaną dzięki modelowi). Wyraża się wzorem [19]:

$$R^2 = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{3.6.}$$

gdzie: $\bar{y}$ – średnia empiryczna wartości rzeczywistych (empirycznych).

W ocenie zasadności zastosowania sieci neuronowej przydatne może być porównanie wyników osiąganych przy użyciu tych struktur, z innymi modelami. W jednym z badań dotyczących ciężkości wypadków zestawiono osiągnięcia sieci neuronowych i innych metod statystycznych, jak regresja nieliniowa, czy prognozowanie z wykorzystaniem rozkładu Poissona. Średni błąd względny MRE wynosił w typ przypadku 34% dla sieci neuronowych, podczas gdy dla innych modeli ok. 44% [14].

Istotną wiedzę może przynieść sprawdzenie wpływu użytych zmiennych na wartość funkcji celu. Działanie to przeprowadza się na koniec, już po zbudowaniu i ocenie modelu. Jednym ze sposobów oceny wpływu różnych czynników na działanie modelu jest zmienianie wartości danej cechy kolejno o -1σ (odchylenie standardowe), +1 σ, +2 σ, +3 σ… Przy każdej zmianie należy obserwować zmianę wartości przewidywanej. Powtarzając procedurę dla kolejnych zmiennych (pamiętając, by pozostałe zmienne w tym czasie miały wartość oryginalną), można uzyskać średnie zmiany wartości przewidywanej wyrażone w procentach. Ich porównanie wskazuje na najbardziej istotne cechy (największa procentowa zmiana wartości przewidywanych) i te mniej znaczące (zmiana wartości w niewielki sposób wpływająca na funkcję celu) [8].

# 4. ZAŁOŻENIA BADAWCZE.

Korzystając z przedstawionych wyżej doświadczeń, przystąpiono do tworzenia własnego modelu bezpieczeństwa ruchu drogowego na odcinkach dróg krajowych w Polsce.

Przyjęto następujące założenia badawczo – projektowe:

- analiza dotyczy tylko zamiejskich, jednojezdniowych odcinków dróg krajowych w Polsce, które były w użyciu w latach 2006 – 2008,

- za odcinek jednojezdniowy uznaje się taki, dla którego na co najmniej 50% długości ruch odbywa się tylko po jednej jezdni. Należy przy tym zaznaczyć, że na ponad 80% odcinków uznanych za jednojezdniowe nie posiada w ogóle drugiej jezdni,

- przy budowie modelu korzystano z gotowej bazy danych obejmującej informacje o wypadkach na odcinkach dróg i cechach tych odcinków,

- zachowano podział dróg na odcinki przyjęty przez autorów bazy danych,

- baza danych uwzględnia informacje o wypadkach, tj. zdarzeniach drogowych, w wyniku których u uczestników nastąpiła śmierć (ofiara zabita) lub uszkodzenie ciała, po którym udzielona została pomoc lekarska (ofiara ranna). Pomija się natomiast informacje o kolizjach drogowych,

- model budowano w oparciu o sztuczne sieci neuronowe z nauczycielem. Oznacza to, że uczenie sieci odbywało się przy znajomości sygnałów wyjściowych (liczby wypadków, ofiar rannych, ofiar ciężko rannych i zabitych),

- na wyjściu modelu starano się uzyskać informacje na temat liczby, gęstości lub koncentracji wypadków. Pominięto natomiast kwestię ciężkości wypadków.

- dane użyte w modelu mają postać zagregowaną, tj. dotyczą łącznej liczby zdarzeń na danym odcinku, nie zawierają natomiast informacji o cechach i przebiegu poszczególnych wypadków,

- przy uczeniu sieci ograniczono się do algorytmu propagacji wstecznej.

# 5. BUDOWA MODELU.

Jednym z celów niniejszej pracy jest zbudowanie modelu, dzięki któremu na podstawie pewnych cech odcinka drogi będzie możliwe określenie charakterystyk bezpieczeństwa ruchu drogowego na tym odcinku. Charakterystyki te obejmują m.in. liczbę, gęstość i koncentrację wypadków. Zadaniem modelu opartego o struktury sztucznych sieci neuronowych będzie więc uogólnianie wiedzy i zdolność do określenia charakterystyk bezpieczeństwa dowolnego odcinka drogi. Ponadto analiza modelu może pozwolić na określenie wpływu poszczególnych cech na poziom bezpieczeństwa.

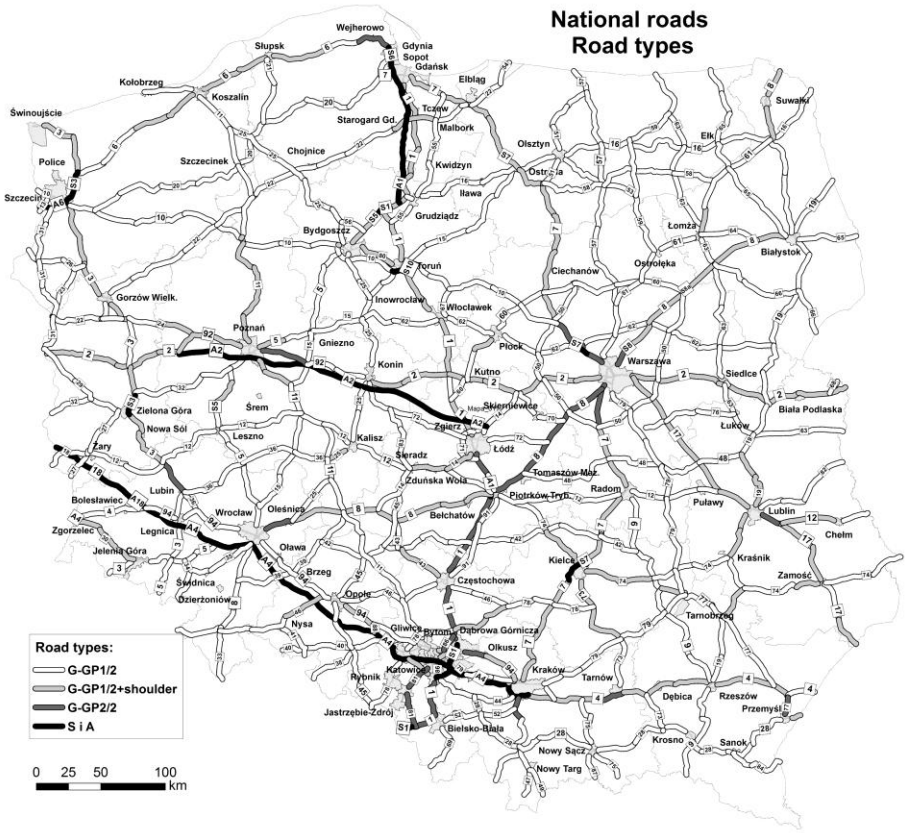
W poniższym rozdziale przedstawiono krok po kroku sposób budowy modelu od przygotowania danych po ocenę ostatecznych efektów jego działania. Opierano się przy tym na wiedzy uzyskanej przez analizę literatury i opisanej w rozdziale 3

## *5.1. Sposób pozyskiwania i przygotowania danych statystycznych.*

Pierwszym krokiem przy budowie każdego modelu jest pozyskanie, a następnie przygotowanie potrzebnych danych. Czynności te w odniesieniu do modelu bezpieczeństwa na drogach krajowych opisano poniżej. Dodatkowo dokonano wstępnej analizy za pomocą jednej z metod grupowania danych zwanej analizą głównych składowych.

### *5.1.1. Źródło i charakterystyka danych.*

Sztuczne sieci neuronowe wymagają do prawidłowego działania dużej ilości danych. Z analizy literatury wynika, że w większości przypadków badacze nie pozyskiwali danych samodzielnie, lecz korzystali z gotowych baz [8,9,11]. Podobnie postąpiono w przypadku niniejszej pracy. Jako źródło danych posłużyła baza stworzona na podstawie SEWiK, czyli Systemu Ewidencji Wypadków i Kolizji, przygotowywanego przez Komendę Główną Policji (*http://www.sewik.pl*). Bazę tę uzyskano bezpośrednio od autorów, którzy korzystali z niej przy własnych badań [6]. Dane obejmują lata 2006 – 2008 i dotyczą sieci wszystkich dróg krajowych w Polsce, które istniały w tym czasie. Mapę tych dróg przedstawiono na rysunku 5.1.




Rysunek 5.1. Mapa dróg krajowych w Polsce w 2008r. z podziałem odcinki i klasy. Źródło: [6]

Zgodnie z założeniami, model bezpieczeństwa ma mieć postać zagregowaną i dotyczyć łącznej liczby wypadków i ofiar na odcinkach dróg. Dlatego też, należało w sposób względnie jednorodny określić te odcinki na sieci dróg krajowych. Również w tej kwestii skorzystano z gotowego rozwiązania, przygotowanego przez autorów „Analizy czynników wpływających na gęstość ofiar śmiertelnych na drogach krajowych w Polsce" [6]. Twórcy analizy dokonali modyfikacji (łączenia) odcinków zaproponowanych w systemie SEWiK. Przyjęto, że w miarę możliwości każdy odcinek powinien cechować się:

- długością przekraczającą 20km
- łączną liczbą ciężko rannych i zabitych w ciągu 3 lat przekraczającą 10 osób.

W przypadku znaczących różnic pomiędzy charakterystykami sąsiednich odcinkami, nie dokonywano łączenia tych odcinków, nawet jeśli powyższe warunki nie zostały spełnione. Ostatecznie, uzyskano 596 odcinków, spośród których część stanowiły odcinki dwujezdniowe, w tym autostrady i drogi szybkiego ruchu [6].

W niniejszej pracy zajęto się wyłącznie pozostałymi odcinkami, czyli tymi uznanymi za jednojezdniowe (udział części odcinka z dwoma jezdniami poniżej 50%). Łączna liczba tych odcinków wyniosła 541. Wartość ta mieści się granicach wielkości zbiorów rekordów, z których korzystali badacze w opisanych w rozdziale 2. modelach wykorzystujących sztuczne sieci neuronowe.

Każdy odcinek w bazie posiada przypisaną nazwę (np. DK01-01 oznacza pierwszy odcinek na drodze krajowej nr 1). Opisany jest zbiorem następujących cech:

a) **DROGA** - Numer drogi.

b) **KD_TYP** - Klasa drogi, gdzie:
- G1 – droga główna jednojezdniowa,
- GP1 – droga główna ruchu przyspieszonego, jednojezdniowa,
- S1 – droga ekspresowa, jednojezdniowa,

c) **KD** – współczynnik [-], przyjęty przez autorów bazy, uwzględniający poziom bezpieczeństwa na poszczególnych klasach drogi:
- dla G1 – 6,5
- dla GP1 – 5
- dla S – 2,2

d) **L** – długość odcinka [km].

e) **LDW** – wskaźnik uwzględniający łączną liczbę wypadków w województwie, na terenie którego położony jest odcinek [-]. Wartość referencyjną 1,000 przyjęto dla województwa Kujawsko – Pomorskiego. Najmniejsza wartość cechowała województwo Małopolskie - 0,557; największa – 1,206 Podlaskie. W przypadku odcinka przebiegającego przez 2 sąsiednie województwo, obliczano współczynnik LDW za pomocą proporcji.

f) **LDZ** – analogicznie, jak w LDW, wskaźnik uwzględniający łączną liczbę zabitych w województwie, na terenie którego położony jest odcinek.

g) **N** – średnioroczne dobowe natężenie ruchu na odcinku [10tys. poj./24h].

h) **NAT** - średnioroczne dobowe natężenie ruchu na odcinku [poj./24h].

i) **UC** – udział pojazdów ciężarowych [-].

j) **PP** – praca przewozowa [mln poj.- km].

g) **LW** – liczba wypadków na odcinku [wyp.].

h) **LR** – liczba rannych [os.].

i) **LCR** – liczba ciężko rannych [os.].

j) **LZ** – liczba zabitych [os.].

k) **Koszt** – koszt wypadków [mln zł].

l) **GW** – gęstość wypadków – stosunek liczby wypadków na danym odcinku do długości tego odcinka [wyp./km/rok].

m) **GR** - gęstość rannych [os./km/rok].

n) **GCR** – gęstość ciężko rannych [os./km/rok].

o) **GZ** – gęstość zabitych [os./km/rok].

p) **GCZ** – gęstość ciężko rannych i zabitych łącznie [os./km/rok].

r) **KW** – koncentracja wypadków – stosunek liczby wypadków na danym odcinku do pracy przewozowej na tym odcinku [wyp./1000 poj.- km].

s) **KR** - koncentracja rannych [os./1000 poj.- km].

t) **KCR** - koncentracja ciężko rannych [os./1000 poj.- km].

u) **KZ** - koncentracja zabitych [os./1000 poj.- km].

w) **KCZ** – koncentracja ciężko rannych i zabitych łącznie [os./1000 poj.- km].

y) **CR** – ciężkość wypadków pod względem liczby rannych – stosunek liczby rannych do liczby wypadków na danym odcinku [os./wyp.].

z) **CZ** - ciężkość wypadków pod względem liczby zabitych – stosunek liczby zabitych do liczby wypadków na danym odcinku [os./wyp.].

aa) **J0** – długość jezdni pierwszej – równa długości odcinka [km].

ab) **J1** – długość jezdni drugiej – w większości przypadków 0 [km].

ac) **P2** – stosunek długości J1 do J0 – w większości przypadków 0, maksymalnie 0,478 [-].

ad) **OZ** – udział długości odcinka prowadzącego w terenie zabudowanym [-].

ae) **PUS** - udział odcinków z szerokim poboczem bitumicznym ≥ 2m [-].

af) **PUW** - udział odcinków z wąskim poboczem bitumicznym < 2m [-].

ag) **PBG** - udział odcinków z poboczem gruntowym [-].

ah) **PAW** - udział odcinków z pasem awaryjnym [-].

ai) **PWL** - udział odcinków z pasem dla ruchu powolnego [-].

aj) **WZ** - gęstość węzłów położonych na danym odcinku [węzłów/km].

ak) **WZZ** - gęstość wjazdów i zjazdów na węzłach [(wj.+zj.)/km].

al) **SKK** - gęstość skrzyżowań z drogami krajowymi [skrz./km].

am) **SKW** - gęstość skrzyżowań z drogami wojewódzkimi [skrz./km].

an) **SKP** - gęstość skrzyżowań z drogami powiatowymi i gminnymi [skrz./km].

ao) **SK** - gęstość skrzyżowań z drogami wszystkich powyższych kategorii [skrz./km].

ap) **SKG** - gęstość skrzyżowań z drogami gruntowymi [skrz./km].

ar) **ZPU** - gęstość zjazdów publicznych [zj./km].

as) **ZPR** - gęstość zjazdów prywatnych [zj./km].

at) **ZLS** - gęstość zjazdów leśnych [zj./km].

au) **CPR** - udział odcinków z ciągiem pieszo - rowerowym [-].

aw) **DR** - udział odcinków zadrzewionych [-].

ay) **GF** - gęstość fotorejestratorów – stosunek liczby fotorejestratorów do długości odcinka [urz./km].

Wybrane cechy zbioru odcinków jednojezdniowych dróg krajowych przedstawiono w tabeli 5.1. Wielkości te pomagają zobrazować skalę wielkości zagadnienia.

Tabela 5.1. Wybrane cechy charakteryzujące badane odcinki dróg za lata 2006 – 2008

| | |
|---|---|
| **Łączna długość badanych odcinków** [km] | 15 142 |
| **Łączna liczba wypadków** [-] | 22 692 |
| **Łączna liczba ofiar rannych** [os.] | 31 430 |
| **Łączna liczba ofiar ciężkorannych** [os.] | 8 340 |
| **Łączna liczba ofiar zabitych** [os.] | 4 142 |
| **Łączna praca przewozowa** [mln poj.-km] | 128,12 |
| **Średnie natężenie dobowe** [poj./24h] | 8 184 |
| **Średni udział pojazdów ciężarowych** [%] | 17,2% |

*5.1.2. Podział i normalizacja danych.*

Gotowe dane poddano odpowiedniemu przygotowaniu, z uwzględnieniem specyfiki i wymagań sztucznych sieci neuronowych. Na początku, dla zróżnicowania rozkładu i wyeliminowania wpływu lokalizacji drogi, posegregowano wszystkie dane w sposób losowy [8,16]. W tym celu posłużono się narzędziami dostępnymi w arkuszu kalkulacyjnym Microsoft Excel.

Korzystając z doświadczeń zagranicznych [8,16], cały zbiór rekordów podzielono w następującym stosunku:
- zbiór uczący – 400 rekordów – 73,9% całości
- zbiór testowy – 141 rekordów – 26,1% całości.

Ponadto, zgodnie z zaleceniami (m.in. [2]) dokonano normalizacji wszystkich danych. Miała ona na celu zniwelowanie dużego zróżnicowania pomiędzy poziomami wielkości wartości poszczególnych cech. Normalizacji dokonano wg formuły 4.1.:

$$X_{ni} = \frac{X_i - \min(X_i)}{\max(X_i) - \min(X_i)} \qquad (5.1.)$$

$X_{ni}$ – wartość znormalizowana cechy X

$X_i$ – wartość cechy przed normalizacją

Normalizacja ta odbywała się już po zaimplementowaniu macierzy wartości w środowisku Scilab za pomocą odpowiedniej komendy. Po przeprowadzeniu procesu uczenia można dokonać denormalizacji, doprowadzając uzyskane wyniki do pierwotnej, bardziej czytelnej postaci. W tym celu przekształcono wzór 5.1., do postaci:

$$Y_i = Y_{ni} * [\max(Y_i) - \min(Y_i)] + \min(Y_i) \quad (5.2.)$$

$Y_{ni}$ – wartość znormalizowana cechy Y uzyskana w prognozie

$Y_i$ – wartość prognozowana cechy Y po denormalizacji

### *5.2. Dobór czynników wyjściowych dla poszczególnych wariantów modelu.*

Bardzo istotnym krokiem przy budowie modelu bezpieczeństwa opartego o sztuczne sieci neuronowe jest określenie oczekiwanych wartości na wyjściach sieci. W tym przypadku stanowią one te spośród cech z bazy danych, które dotyczą wypadków i ich skutków. Dla przewidywania opisanych poniżej zestawów cech wyjściowych, budowano modele, różniące się typami i ilością wejść oraz strukturą. Następnie oceniano przydatność tych modeli pod kątem estymacji różnych zmiennych celu.

Dostępny zestaw danych pozwala na budową modelu zagregowanego [20]. W związku z tym, jako zmienne celu przyjmowano zestawy cech odpowiednie dla tego typu modeli [8,9]. Zasadniczo, każdy zestaw składał się z czterech cech, prognozowanych jednocześnie przez sieć o danej strukturze. Jest to zgodne z wytycznymi przedstawionymi w literaturze [10].

Ostatecznie, jako wartości wyjściowe zastosowano następujące zestawy liczby i zestawy cech:

I. 4 wyjścia – LW, LR, LCR, LZ. Są to wartości bezwzględne, silnie zależne od długości danego odcinka. Z drugiej strony stanowią najprostszy i najczytelniejszy wskaźnik mogący określać poziom bezpieczeństwa.

II. 4 wyjścia – GW, GR, GCR, GZ. Gęstości wypadków, rannych, ciężko rannych i zabitych pozwalają ocenić rzeczywisty poziom bezpieczeństwa odcinka bez względu na jego długość. Podstawowy wpływ będą miały w tym przypadku pozostałe cechy odcinka. Gęstości wyrażane są w liczbie wypadków na 1

kilometr [wyp./km] lub liczbie ofiar rannych, ciężko rannych, śmiertelnych na kilometr [os./km].

III. 4 wyjścia – KW, KR, KCR, KZ. Koncentracje odnoszą liczbę wypadków i ofiar do pracy przewozowej na danym odcinku, co pozwala ograniczyć wpływ zarówno długości, jak i natężenia ruchu na danym odcinku (praca przewozowa jest iloczynem tych wielkości). W związku z tym, możliwe jest najbardziej istotnego znaczenia nabierają cechy charakteryzujące odcinek, nie związane z ruchem. Koncentracje wyrażone są w liczbie wypadków lub ofiar (rannych, ciężko rannych, śmiertelnych) na 1000 pojazdo-kilometrów [wyp./1000 poj.-km; os./1000 poj.-km].

Ponadto, w celu porównania możliwości sztucznych sieci neuronowych budowano również modele zawierające na wyjściu tylko jedną z cech. Korzystano przy tym z tych samych zmiennych, co w powyżej opisanych przypadkach. Porównanie wyników może dać odpowiedź na pytanie o zasadność osobnego budowania modeli dla przewidywania różnych cech. Potencjalnie, istnieje możliwość, że model będzie w stanie bardziej dokładnie przewidywać jedną cechę, niż cały ich zestaw na raz.

Do budowania tego rodzaju modeli wybrano następujące cechy, uznane za podstawowe wskaźniki bezpieczeństwa:

IV. 1 wyjście – LW
V. 1 wyjście – LZ
VI. 1 wyjście – LCR
VII. 1 wyjście – GW
VIII. 1 wyjście – GZ
IX. 1 wyjście – KW
X. 1 wyjście – KZ

Należy zaznaczyć, że przy przeprowadzaniu testów kolejnych modeli z jednym wyjściem, kierowano się w pewnej mierze rezultatami uzyskiwanymi w modelach z czterema wyjściami. Oznacza to, iż zmienne wyjściowe dające najbardziej obiecujące wyniki spośród grupy podobnych czynników (np. liczb bezwzględnych) poddawane były bardziej szczegółowej analizie, niż te dla których prognozy okazywały się mniej dokładne.

W porównaniu z dotychczas stworzonymi modelami główna różnica polega na zastosowaniu dodatkowego wskaźnika, jakim jest koncentracja (wypadków, rannych, ciężko rannych lub zabitych). Z drugiej strony, skorzystano również z najczęściej spotykanych sposobów oceny bezpieczeństwa, tj. wartości bezwzględnych dotyczących liczb wypadków i ofiar oraz gęstości tychże. Pominięto natomiast kwestię ciężkości wypadków, aczkolwiek możliwe byłoby zbudowanie na podstawie dostępnych danych modelu dotyczącego również tego problemu. Należałoby wówczas za cechę wyjściową uznać CR lub CZ.

### *5.3. Dobór czynników wejściowych dla poszczególnych wariantów modelu.*

Równie istotnym zagadnieniem, co dobór wyjść modelu jest dobór jego cech wejściowych, czyli objaśniających. Zmienne te stanowią dla sztucznych sieci neuronowych podstawę do estymacji funkcji celu (zmiennych wyjściowych).

#### *5.3.1. Przyjęte kryteria doboru.*

Z analizy prac zagranicznych dotyczących tworzenia modeli bezpieczeństwa na odcinkach dróg wynika, że w większości przypadków badacze w wyborze czynników wejściowych kierowali się wyłącznie wcześniejszymi doświadczeniami [9,10,13]. W niektórych przypadkach zastosowano pewne metody matematyczne [12,13,14]. Jedną z metod przygotowania i klasyfikacji danych, która jednocześnie może posłużyć do wstępnej selekcji zmiennych, jest metoda PCA, opisana w podpunkcie 3.3.3.

Mając powyższe na uwadze, zdecydowano się na dobranie zestawów zmiennych wejściowych w oparciu o analizę przedstawionych i opisanych w rozdziale 3. przykładów podobnych modeli. Ponadto, skorzystano również z metody PCA, jako dodatkowej możliwości analizy problemu i ustalenia najważniejszych czynników, które powinny zostać uwzględnione przy predykcji wypadków.

Przy doborze wejść modelu dla polskich dróg krajowych należało jednak zwrócić uwagę także na dostępność danych. Niestety, wielu z czynników uznawanych za istotne, nie uwzględniono w bazie danych, z której zamierzano korzystać. Przede wszystkim, są tą: dozwolona prędkość i szerokość drogi na danym odcinku [9,10], a także cechy wypadków i uczestników (potrzebne w przypadku budowy modelu nieskumulowanego) [13].

*5.3.2. Analiza Głównych Składowych (PCA).*

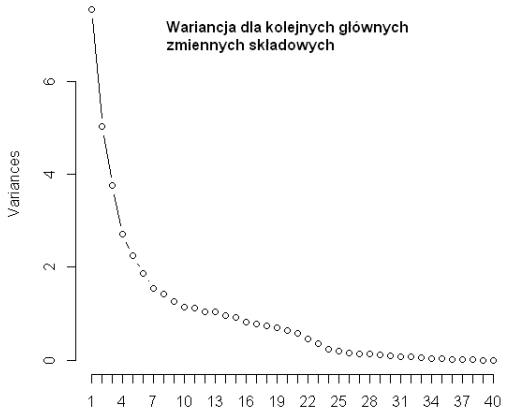
Analizy PCA dokonano w oparciu o 28 zmienne charakteryzujące odcinek drogi i ruch na nim oraz 12 wskaźników bezpieczeństwa opisanych w punkcie 5.2. Łącznie analizie poddano 40 zmiennych, zarówno wyjściowych, jak i wejściowych dla sieci. Jako narzędzie posłużył interfejs RGUI, oparty o język programowania R. Po wczytaniu zmiennych dokonano ich normalizacji, zgodnie z zaleceniami Gramackiego [18]. Sama Analiza Głównych Składowych przeprowadzana jest w oparciu o gotową funkcję i nie wymaga dodatkowych czynności. Wyniki analizy przedstawiono za równo w postaci tabelarycznej, jak i za pomocą podstawowych wykresów opisanych w podpunkcie 3.3.3.

Z tabeli 5.2., będącej efektem przeprowadzonej analizy PCA wynika, że pierwsza główna składowa (PC1) tłumaczy w ok. 19% zmienność całego zbioru i najbardziej wpływa na odchylenie standardowe. Wzięcie pod uwagę 10 pierwszych zmiennych spowoduje utratę niespełna 29% informacji. Do dalszych analiz warto pozostawić jednak 15 pierwszych zmiennych głównych składowych, których uwzględnienie pozwoli zachować 84% oryginalnej zmienności zbioru. Pozostałe zmienne można odrzucić, gdyż w niewielkim stopniu tłumaczą one zmienność zbioru. Redukcja ilości danych ułatwia ich analizę, może też dostarczyć więcej informacji, niż kompletny zbiór [18].

Tabela 5.1. Analiza PCA – podsumowanie wariancji objaśnianych przez kolejne główne składowych

| | **PC1** | **PC2** | **PC3** | **PC4** | **PC5** | **PC6** | **PC7** |
|---|---|---|---|---|---|---|---|
| Odchylenie standardowe | 2,7469 | 2,2415 | 1,93849 | 1,6451 | 1,49815 | 1,3639 | 1,23967 |
| Odsetek wyjaśniający wariancję | 0,1886 | 0,1256 | 0,09394 | 0,06766 | 0,05611 | 0,0465 | 0,03842 |
| Skumulowany odsetek | 0,1886 | 0,3142 | 0,40819 | 0,47585 | 0,53196 | 0,5785 | 0,61689 |
| | **PC8** | **PC9** | **PC10** | **PC11** | **PC12** | **PC13** | **PC14** |
| Odchylenie standardowe | 1,19448 | 1,1257 | 1,0687 | 1,05754 | 1,02293 | 1,01891 | 0,98201 |
| Odsetek wyjaśniający wariancję | 0,03567 | 0,03168 | 0,02855 | 0,02796 | 0,02616 | 0,02595 | 0,02411 |
| Skumulowany odsetek | 0,65256 | 0,68424 | 0,71279 | 0,74075 | 0,76691 | 0,79286 | 0,81697 |
| | **PC15** | **PC16** | **PC17** | **PC18** | **PC19** | **PC20** | **PC21** |
| Odchylenie standardowe | 0,96363 | 0,9078 | 0,88357 | 0,85925 | 0,83741 | 0,79737 | 0,76333 |
| Odsetek wyjaśniający wariancję | 0,02321 | 0,0206 | 0,01952 | 0,01846 | 0,01753 | 0,01589 | 0,01457 |
| Skumulowany odsetek | 0,84019 | 0,8608 | 0,88031 | 0,89877 | 0,9163 | 0,93219 | 0,94676 |
| | **PC22** | **PC23** | **PC24** | **PC25** | **PC26** | **PC27** | **PC28** |
| Odchylenie standardowe | 0,67456 | 0,6027 | 0,4764 | 0,43223 | 0,38623 | 0,3743 | 0,35129 |
| Odsetek wyjaśniający wariancję | 0,01138 | 0,00908 | 0,00567 | 0,00467 | 0,00373 | 0,0035 | 0,00309 |
| Skumulowany odsetek | 0,95813 | 0,96722 | 0,97289 | 0,97756 | 0,98129 | 0,9848 | 0,98788 |
| | **PC29** | **PC30** | **PC31** | **PC32** | **PC33** | **PC34** | **PC35** |
| Odchylenie standardowe | 0,3272 | 0,29431 | 0,27063 | 0,26443 | 0,23456 | 0,18407 | 0,1694 |
| Odsetek wyjaśniający wariancję | 0,00268 | 0,00217 | 0,00183 | 0,00175 | 0,00138 | 0,00085 | 0,00072 |
| Skumulowany odsetek | 0,99055 | 0,99272 | 0,99455 | 0,9963 | 0,99767 | 0,99852 | 0,99924 |
| | **PC36** | **PC37** | **PC38** | **PC39** | **PC40** | | |
| Odchylenie standardowe | 0,13751 | 0,09404 | 0,05242 | 0,001851 | 0,000492 | | |
| Odsetek wyjaśniający wariancję | 0,00047 | 0,00022 | 0,00007 | 0 | 0 | | |
| Skumulowany odsetek | 0,99971 | 0,99993 | 1 | 1 | 1 | | |

Potwierdzeniem wniosków wyciągniętych z analizy tabeli, jest wykres 5.2. Na osi odciętych zaznaczono kolejne numery głównych składowych, a na osi rzędnych wielkość wariancji zbioru. Widać na nim wyraźnie, że pierwsze 5 - 6 głównych składowych w dużo większym stopniu decyduje o wariancji zbioru, niż pozostałe.




Rysunek 5.2. Analiza PCA – wariancja zbioru objaśniana przez poszczególne główne składowe.

Oprócz redukcji wymiarowości zbioru, metoda PCA dostarcza informacji o istotności rzeczywistych zmiennych. Z wykresu zmiennych (biplot) – rysunek 5.3. – można odczytać wpływ zmiennych na dwie pierwsze składowe główne, poprzez porównanie długości, kierunku i zwrotu wektorów poszczególnych zmiennych. Ponadto na wykresie zawarto numery reprezentujące kolejne odcinki dróg. Numery te traktowane są jako punkty w przestrzeni wielowymiarowej zrzutowane na płaszczyznę wyznaczoną przez PC1 i PC2 Ich położenie wskazuje na ich wzajemną zależność, odchylenia od normy, a także związek ze zmiennych objaśniającymi. Z analizy wykresu można wyciągnąć następujące wnioski:

- występuje silna zależność liczby zabitych (LZ) i gęstości zabitych (GZ) od natężenia ruchu (NAT) i pracy przewozowej (PP),

- takie zmienne, jak gęstości węzłów (WZ), zjazdów na węzłach (WZZ), czy szczególnie wskaźniki oznaczające województwa (LDZ, LDW) są niemal prostopadłe do wymienionych wskaźników bezpieczeństwa (LZ, GZ), a więc cechują się niewielką współzależnością,

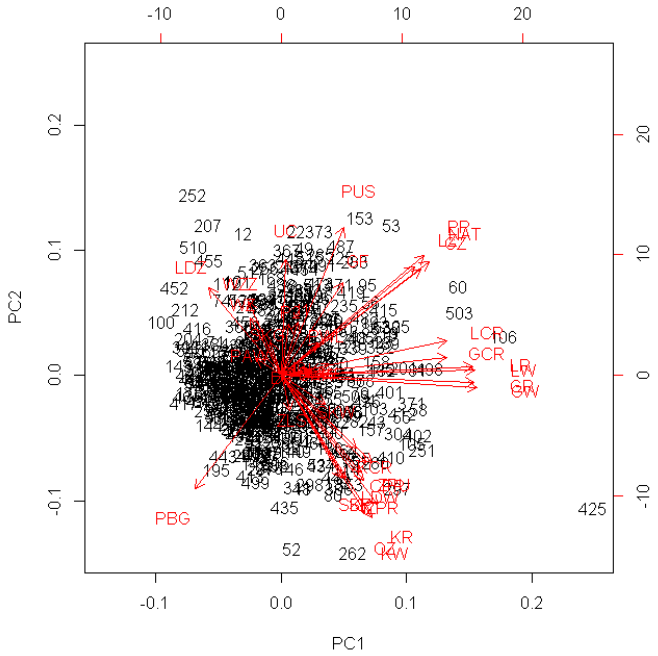
- zmienne WZ, WZZ, LDZ, LDW są skierowane przeciwnie do m.in. koncentracji wypadków (KW), czy koncentracji rannych (KZ), co wskazuje na ich ujemną korelację. Oznacza to, że wzrost gęstości węzłów i zjazdów na węzłach powoduje spadek koncentracji wypadków i ofiar śmiertelnych,

- dodatnią korelację z koncentracjami posiadają natomiast wskaźniki dotyczące liczby skrzyżowań. Wobec tego, wzrost liczby punktów kolizji skutkuje większą koncentracją zdarzeń,

- jedyną wyraźnie odstającą od reszty zmienną jest PBG, a więc udział odcinków z poboczem gruntowym. Kierunek i zwrot jej wektora wskazuje, że duży udział takich odcinków wiąże się ze zmniejszeniem wskaźników bezpieczeństwa, takich jak LZ, GZ, czyli poprawą bezpieczeństwa. Może to być związane z faktem, że takie pobocza rzadziej występują na drogach o dużym natężeniu ruchu i pracy przewozowej,

- wektory wymienionych wyżej zmiennych, wraz z innymi wskaźnikami dotyczącymi liczby i gęstości wypadków oraz kilkoma cechami odcinków posiadają największą długość, a więc największy wpływ na PC1 i PC2.

- spośród opisanych odcinków dróg wyraźnie odstaje od reszty odcinek nr 425 (DK35-02).




Rysunek 5.3. Analiza PCA - wykres wpływu zmiennych na składowe główne

Potwierdzenie wniosków z analizy wykresu można znaleźć w tabeli 5.3. Poszczególne wskaźniki oznaczają wagi przypisane poszczególnym zmiennym. Wyliczenie kombinacji liniowej przy użyciu tychże zmiennych i wag, pozwala zastąpić je składową główną. Jest to potrzebne w przypadku chęci zastosowania składowych w dalszej analizie, np. w sieciach neuronowych. Dodatkowo wielkości i znaki informują o znaczeniu poszczególnych zmiennych. Zwraca przy tym uwagę znikomy wpływ udziału pojazdów ciężkich na najważniejszą składową PC1.

Tabela 5.2. Analiza PCA – wskaźniki dla 15 najistotniejszych składowych głównych.

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 | PC10 | PC11 | PC12 | PC13 | PC14 | PC15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **LW** | 0,314 | 0,010 | -0,155 | 0,057 | -0,148 | 0,003 | -0,119 | 0,007 | -0,052 | 0,101 | -0,092 | -0,027 | -0,025 | 0,001 | 0,025 |
| **LR** | 0,310 | 0,018 | -0,152 | 0,023 | -0,144 | 0,012 | -0,160 | -0,027 | -0,072 | 0,100 | -0,098 | -0,021 | -0,048 | -0,002 | 0,045 |
| **LCR** | 0,268 | 0,069 | -0,135 | -0,078 | -0,053 | 0,156 | -0,029 | -0,046 | 0,260 | -0,253 | 0,039 | -0,045 | 0,146 | -0,017 | -0,050 |
| **LZ** | 0,216 | 0,216 | -0,204 | 0,067 | -0,023 | -0,142 | 0,243 | -0,002 | -0,039 | 0,085 | 0,086 | -0,028 | 0,119 | -0,050 | 0,001 |
| **GW** | 0,317 | -0,023 | 0,100 | -0,128 | 0,078 | 0,063 | -0,061 | 0,076 | -0,092 | 0,167 | -0,110 | 0,015 | -0,045 | 0,007 | 0,013 |
| **GR** | 0,313 | -0,015 | 0,082 | -0,150 | 0,072 | 0,070 | -0,104 | 0,034 | -0,121 | 0,154 | -0,108 | 0,031 | -0,065 | 0,007 | 0,032 |
| **GCR** | 0,268 | 0,037 | 0,053 | -0,224 | 0,124 | 0,194 | -0,008 | 0,009 | 0,204 | -0,187 | 0,018 | 0,003 | 0,102 | -0,014 | -0,051 |
| **GZ** | 0,227 | 0,211 | -0,007 | -0,079 | 0,108 | -0,162 | 0,320 | 0,049 | -0,080 | 0,135 | 0,030 | -0,004 | 0,140 | -0,059 | 0,030 |
| **KW** | 0,147 | -0,282 | -0,117 | -0,272 | -0,070 | -0,051 | 0,070 | -0,049 | -0,104 | 0,021 | -0,018 | -0,018 | -0,156 | 0,029 | 0,043 |
| **KR** | 0,156 | -0,256 | -0,124 | -0,290 | -0,068 | -0,021 | -0,004 | -0,096 | -0,121 | 0,015 | -0,036 | -0,005 | -0,182 | 0,044 | 0,059 |
| **KCR** | 0,121 | -0,144 | -0,103 | -0,337 | 0,023 | 0,154 | 0,087 | -0,115 | 0,267 | -0,380 | 0,112 | -0,038 | 0,048 | 0,037 | -0,038 |
| **KZ** | 0,049 | 0,011 | -0,198 | -0,197 | 0,001 | -0,255 | 0,545 | -0,027 | -0,065 | 0,085 | 0,178 | -0,003 | 0,063 | -0,095 | 0,031 |
| **L** | 0,022 | 0,013 | -0,403 | 0,231 | -0,274 | 0,027 | -0,054 | -0,061 | 0,071 | -0,032 | -0,026 | -0,034 | -0,004 | 0,051 | -0,039 |
| **LDW** | 0,128 | -0,193 | 0,115 | -0,062 | -0,158 | -0,368 | -0,193 | -0,065 | -0,125 | -0,031 | 0,180 | 0,015 | 0,061 | 0,086 | -0,015 |
| **LDZ** | -0,117 | 0,174 | -0,135 | 0,022 | 0,168 | 0,324 | 0,325 | 0,149 | 0,079 | 0,009 | -0,123 | -0,049 | 0,004 | 0,027 | -0,021 |
| **NAT** | 0,239 | 0,227 | 0,194 | 0,066 | 0,116 | 0,071 | -0,067 | 0,078 | -0,055 | 0,103 | -0,064 | 0,054 | 0,028 | -0,021 | 0,023 |
| **UC** | 0,008 | 0,231 | -0,094 | 0,018 | 0,125 | -0,091 | 0,139 | 0,048 | -0,083 | -0,255 | -0,039 | -0,101 | -0,115 | 0,122 | 0,302 |
| **PP** | 0,232 | 0,239 | -0,110 | 0,237 | -0,084 | 0,049 | -0,115 | 0,001 | 0,007 | 0,052 | -0,047 | 0,004 | 0,035 | -0,020 | 0,008 |
| **J0** | 0,022 | 0,013 | -0,403 | 0,231 | -0,274 | 0,027 | -0,054 | -0,061 | 0,071 | -0,033 | -0,026 | -0,034 | -0,004 | 0,051 | -0,039 |
| **J1** | 0,027 | 0,105 | 0,182 | 0,000 | -0,404 | 0,318 | 0,139 | -0,026 | -0,181 | -0,053 | 0,241 | 0,215 | -0,049 | -0,062 | 0,154 |
| **P2** | 0,012 | 0,102 | 0,241 | -0,028 | -0,386 | 0,288 | 0,145 | -0,009 | -0,212 | -0,070 | 0,219 | 0,184 | -0,057 | -0,061 | 0,087 |
| **OZ** | 0,136 | -0,273 | 0,117 | 0,200 | -0,041 | 0,049 | 0,120 | 0,267 | -0,033 | -0,113 | -0,033 | -0,051 | 0,091 | 0,068 | 0,028 |
| **PUS** | 0,102 | 0,294 | 0,112 | 0,018 | 0,115 | -0,157 | -0,067 | -0,102 | 0,004 | -0,116 | 0,077 | -0,032 | -0,183 | 0,175 | -0,063 |
| **PUW** | 0,049 | 0,005 | 0,067 | 0,035 | -0,125 | -0,286 | -0,012 | 0,154 | 0,315 | -0,100 | 0,182 | 0,122 | -0,309 | -0,485 | -0,103 |
| **PBG** | -0,140 | -0,225 | -0,211 | -0,089 | -0,050 | 0,217 | 0,050 | 0,046 | -0,141 | 0,132 | -0,122 | 0,028 | 0,274 | 0,055 | 0,100 |
| **PAW** | -0,041 | 0,033 | 0,079 | -0,074 | -0,088 | -0,071 | -0,127 | -0,048 | -0,111 | 0,187 | 0,168 | -0,354 | 0,566 | -0,367 | 0,021 |
| **PWL** | 0,060 | 0,065 | 0,040 | -0,027 | -0,005 | -0,062 | -0,033 | 0,080 | 0,102 | 0,103 | 0,321 | 0,312 | 0,310 | 0,548 | -0,404 |
| **WZ** | -0,050 | 0,114 | 0,175 | -0,140 | -0,294 | -0,211 | 0,116 | 0,009 | 0,159 | 0,008 | -0,513 | 0,001 | 0,105 | 0,168 | 0,041 |
| **WZZ** | -0,053 | 0,148 | 0,204 | -0,137 | -0,382 | -0,145 | 0,136 | 0,007 | 0,135 | -0,004 | -0,358 | -0,016 | 0,017 | 0,101 | -0,034 |
| **SKK** | 0,047 | 0,003 | 0,064 | 0,062 | 0,069 | 0,040 | 0,009 | -0,134 | -0,503 | -0,416 | -0,171 | -0,119 | 0,028 | -0,023 | -0,255 |
| **SKW** | 0,072 | -0,057 | 0,089 | 0,056 | 0,040 | 0,217 | 0,096 | -0,053 | 0,201 | 0,456 | 0,135 | -0,243 | -0,307 | 0,131 | 0,065 |
| **SKP** | 0,097 | -0,203 | 0,133 | 0,280 | 0,061 | -0,077 | 0,209 | -0,428 | 0,047 | -0,001 | -0,021 | 0,054 | 0,045 | 0,081 | 0,093 |
| **SK** | 0,103 | -0,204 | 0,140 | 0,281 | 0,066 | -0,056 | 0,214 | -0,429 | 0,043 | 0,021 | -0,016 | 0,028 | 0,020 | 0,089 | 0,086 |
| **SKG** | -0,017 | 0,067 | 0,081 | -0,017 | -0,169 | 0,058 | 0,118 | -0,063 | -0,098 | 0,009 | 0,113 | -0,623 | -0,208 | 0,096 | -0,494 |
| **ZPU** | 0,147 | -0,173 | 0,114 | 0,083 | 0,074 | 0,106 | 0,089 | 0,180 | 0,074 | 0,017 | -0,248 | -0,025 | 0,091 | -0,166 | -0,197 |
| **ZPR** | 0,133 | -0,208 | 0,061 | 0,222 | -0,078 | -0,105 | 0,109 | 0,282 | 0,085 | -0,169 | 0,042 | -0,062 | 0,092 | -0,020 | 0,033 |
| **ZLS** | 0,015 | -0,070 | -0,106 | 0,080 | 0,038 | -0,167 | 0,065 | 0,468 | -0,286 | -0,066 | 0,084 | -0,016 | -0,146 | 0,220 | 0,093 |
| **CPR** | 0,138 | -0,174 | 0,171 | 0,271 | -0,003 | 0,078 | 0,132 | 0,193 | 0,106 | -0,066 | 0,000 | -0,054 | 0,058 | -0,037 | -0,066 |
| **DR** | 0,003 | -0,002 | -0,141 | 0,046 | 0,084 | 0,015 | 0,127 | -0,117 | -0,172 | 0,106 | -0,178 | 0,423 | -0,105 | -0,274 | -0,496 |
| **GF** | 0,100 | 0,184 | 0,026 | 0,101 | 0,135 | -0,009 | -0,087 | -0,168 | -0,057 | -0,179 | 0,018 | -0,039 | 0,068 | -0,048 | 0,191 |

Podsumowując, analiza PCA przyniosła szereg informacji dotyczących wyboru i znaczenia zmiennych. Wnioski z tej analizy mogą pomóc w budowie i lepszym zrozumieniu modelu opartego o sztuczne sieci neuronowe.

### *5.3.3. Przyjęte zestawy czynników wejściowych.*

W celu zbudowania możliwie najlepszego modelu i porównania różnych jego wersji oraz wykorzystania możliwości, jakie daje dostępna baza, w niniejszej pracy poddawano ocenie modele wykorzystujące różne zestawy cech. Co więcej, przewidywano, że różne zestawy zmiennych objaśniających mogą okazać się pomocne przy prognozowaniu różnych zestawów cech wyjściowych, opisanych w punkcie 4.2.

Ostatecznie, modele budowano w oparciu o następujące kombinacje cech podawanych na wejściu sieci:

A) Kompletny zestaw 28 zmiennych opisujących odcinek. Znalazły się tutaj wszystkie cechy odcinków, z wyjątkiem tych dotyczących samych wypadków, ponieważ zostały one użyte, jako wyjścia sieci. Ponadto, pominięto kolumny DROGA i KD_TYP, będące wartościami tekstowymi oraz N, która jest odpowiednikiem NAT, wyrażonym w innych jednostkach. Tak więc w modelu tym skorzystano z: L, LDW, LDZ, NAT, UC, PP, J0, J1, P2, OZ, PUS, PUW, PBG, PAW, PWL, WZ, WZZ, SKK, SKW, SKP, SK, SKG, ZPU, ZPR, ZLS, CPR, DR, GF.

B) Wybrane 22 cechy opisujące odcinek. W wariancie tym dokonano odrzucenia sześciu cech w stosunku do podpunktu A). Po pierwsze, jest to L – długość odcinka, ze względu na jej przewidywany duży i oczywisty wpływ na liczbę wypadków. L mogło zyskiwać dominujące znaczenie przy procesie uczenia zwłaszcza przy LW, LR, LCR i LZ, jako wyjściach sieci. Ponadto dość arbitralny charakter podziału na odcinki sprawia, że uwzględnienie L może nie przynosić konstruktywnych wniosków. Poza tym pominięto LDW i LDZ, jako związane dane związane już z liczbą wypadków. Ponadto usunięto J0 i J1, jako że wskaźnik P2 wyraża te cechy w wystarczającym stopniu. Ostatnią z usuniętych cech jest SKG, z uwagi na fakt, iż tylko w przypadku kilku rekordów posiada wartość różną od 0.

C) Model ograniczony do 14 cech opisujących odcinek. W stosunku do podpunktu B), usunięto kolejne kilka zmiennych. Były to PAW i PWL, które w odpowiednio

98 i 91% miały wartość 0. Z dwóch kolumn dotyczących gęstości węzłów zachowano jedną, usuwając WZZ. Podobnie w przypadku skrzyżowań, zrezygnowano z rozgraniczania dróg pod względem kategorii, eliminując SKK, SKW, SKP. Również gęstości zjazdów ZPU, ZPR, ZLS zastąpiono syntetycznym wskaźnikiem **Z**, będącym sumą gęstości zjazdów publicznych, prywatnych i leśnych.

D) Model podstawowy zawierający 3 najważniejsze cechy. W modelu tym uwzględniono wyłącznie podstawowe cechy, uznawane powszechnie za najbardziej wpływające na liczbę wypadków. Wszystkie one dotyczą cech ruchu, a więc są to NAT, UC oraz PP.

E) Model dodatkowy, zawierający 12 cech opisujących odcinek. Są to wszystkie zmienne wymienione w podpunkcie C), za wyjątkiem uznawanych zwykle za bardzo istotne, pracy przewozowej i natężenia pojazdów. Model ten może przynieść wiedzę na temat wpływu pozostałych cech na bezpieczeństwo i ich znaczenia dla powstawania wypadków.

Warto zauważyć, że wszystkie wymienione powyżej zestawy cech różnią się również co do ilości zmiennych. Fakt ten może mieć istotne znaczenie w kwestii osiąganych wyników. Możliwość uproszczenia struktury sztucznej sieci neuronowej przez zmniejszenie liczby wejść jest cenna ze względu na przyspieszenie procesu uczenia, ale może również pomóc w osiąganiu lepszych rezultatów [15]. Bez wątpienia przy braku istotnych różnic w dokładności wyników, za korzystniejszą należałoby uznać strukturę prostszą.

Podsumowując, w porównaniu z przykładami opisanymi w rozdziale 2., uzyskano pewną różnorodność, dobierając kilka odmiennych zestawów cech wejściowych. Możliwe będzie dzięki temu znalezienie optymalnej liczby i kombinacji zmiennych do stworzenia jak najdokładniejszego modelu. Z podobieństw z przykładami z literatury, należy wymienić uwzględnienie dodatkowych cech drogi, takich jak występowania pobocza, czy liczba skrzyżowań oraz podstawowych cech dotyczących ruchu.

## *5.4. Struktura i sposoby oceny modelu.*

Po dobraniu wejść i wyjść dla sztucznej sieci neuronowej można było przystąpić do jej budowy i testowania poszczególnych wariantów przy różnej architekturze sieci. Aby testowanie było możliwe, należało wcześniej określić kryteria oceny dokładności modelu.

### *5.4.1. Narzędzia i metodyka modelowania.*

Do budowy modeli o różnej strukturze sieci neuronowej użyto pakietu Scilab i Toolboxa ANN [23], które stanowiły podstawowe narzędzie niniejszej pracy. W oparciu o nie, napisano skrypt (kod program w załączniku), w wyniku którego były wykonywane następujące czynności:

1) Wczytanie danych z pliku i ich normalizacja (sposób normalizacji przedstawiono we wzorze 5.1.).
2) Określenie liczby warstw i neuronów w warstwach.
3) Określenie wartości pozostałych cech sieci.
4) Określenie liczby epok w procesie uczenia.
5) Uzyskanie wyników i ich ocena dla zbioru uczącego.
6) Uzyskanie wyników i ich ocena dla zbioru testowego.
7) Wykonanie wykresu dopasowania wartości wyznaczonych przez sieć z rzeczywistymi

### *5.4.2. Elementy struktury modelu.*

Z punktu widzenia określenia struktury modelu najistotniejsze są punkty 2-4. W kolejnych wariantach modelu dokonywano w nich szeregu zmian. Ich celem było uzyskanie dla poszczególnych kombinacji warunków wyjściowych i wejściowych optymalnej struktury sztucznej sieci neuronowej. Struktura taka miała pozwolić jak najdokładniej dokonywać prognozować oczekiwane wartości wyjściowych na podstawie zmiennych wejściowych.

Jeśli chodzi o **liczbę warstw**, testowano przede wszystkim modele z jedną warstwą ukrytą, w mniejszym stopniu z dwiema i trzema warstwami ukrytymi, zgodnie z przyjmowanymi przez większość badaczy założeniami [9,10,13]. Liczbę neuronów w tej części sieci zmieniano zwykle w granicach od 5 do 20 przy jednej warstwie ukrytej oraz od 2 do 20 przy 2 lub trzech warstwach ukrytych. Przedziały te zawierają szacunkowe

wartości wyliczone wg wzorów 3.1 i 3.2 dla poszczególnych układów wejść i wyjść. Przy tym liczba neuronów na wejściu jest określona na podstawie liczby cech wejściowych, które dany model ma uwzględniać. Na wyjściu liczba ta wynosi 1 lub 4 neurony, w zależności od tego, czy model ma jednocześnie prognozować każdy ze wskaźników z osobna, czy łącznie. W odniesieniu do opisanych w podpunkcie 2.3.4. przykładów, tworzony model nie różnił się znacząco pod względem liczby warstw i neuronów.

Toolbox ANN daje szereg możliwości, jeśli chodzi o ustalenie szczegółów charakterystyki sztucznej sieci neuronowej [23]. Wszystkie struktury budowane w programie są sieciami z nauczycielem (feedforward net), jednak dostępne są różne sposobu uczenia. W niniejszych badaniach ograniczono się do podstawowego, stosowanego również w innych modelach bezpieczeństwa na odcinkach dróg [9,10], tj. propagacja wsteczna (online backpropagation with momentum). Zastosowanie tego algorytmu wymaga ustalenia czterech parametrów uczenia (lp), jako, jednego z argumentów funkcji ann_FF_Mom_online.

Każdorazowo należało więc ustalić następujące parametry uczenia:

- **Learning rate** - parametr dotyczący szybkości uczenia. Jego zakres wynosi od 0 do 1, przy czym typowo przyjmuje się wartości między 0,05, a 0,75. Przyjęcie niskiego współczynnika może spowolnić proces uczenia, natomiast zbyt współczynnik uczenia się może powodować uzyskiwanie nieoptymalnych rozwiązań. Wpływ współczynnika może być w praktyce niwelowany przez zmianę liczby epok uczenia, stąd podczas badań dokonywano zmian współczynnika szybkości uczenia w niewielkim stopniu.

- **Tolerated error** - parametr dotyczący błędu pomijalnego. Typowy zakres tego parametru wynosi od 0 do 0,1. Współczynnik ten określa wielkość błędu, który przez sieć zostanie uznany za pomijalny i który nie musi być poprawiany w kolejnych epokach. Należy zauważyć, że parametr ten nie jest tożsamy z poziomem dopasowania $R^2$ i nie może być traktowany, jako oczekiwany poziom dokładności. W poszczególnych modelach przyjmowano różne wartości współczynnika, w sugerowanym zakresie.

- **Momentum** („pęd"). Zakres tego parametru również wynosi od 0 do 1. Jest to jeden z podstawowych parametrów dla sieci z propagacją wsteczną. Określa on poziom stabilizacji zmian wag w procesie uczenia. Właściwe dobranie momentum pozwala przyspieszyć proces uczenia i uniknąć osiągania minimów lokalnych w

poszukiwaniu najlepszych rozwiązań [21]. Na potrzeby poszukiwania optymalnych rozwiązań, podczas badań przyjmowano zwykle wartości 0,0 bądź 0,5.

- **Flat spot elimination constant** - stała eliminacji błędu flat spot (błąd „płaskich pól"). W przypadku tego parametru można przyjmować wartości od 0 do 0,25. Zwiększanie wartości stałej pozwala unikać sytuacji, w której następuje przerwania ze względu na niewielkie zmiany w poziomie poprawy błędu. W poszczególnych modelach przyjmowano wartości parametru 0, 0,1 lub 0,25.

**Liczbę epok** uczenia w poszczególnych wariantach różnicowano w przedziale od 50 do 1000. W wielu przykładach modeli, liczba epok nie była z góry ustalona, a sieć uczono tak długo, aż osiągnięto pożądane wartości błędów, monitorowane na bieżąco [9]. W niniejszych badaniach również monitorowano wartość błędu na zbiorze uczącym co 50 epok. Łączną liczbę epok określano odgórnie, przed rozpoczęciem nauki sieci. Do optymalnej długości procesu uczenia dochodzono metodą prób i błędów dla każdego z przypadków z osobna. Korzystano przy tym również z analizy monitorowanych wartości błędów w zbiorze uczącym.

Podsumowując, dla różnych kombinacji liczby oraz rodzajów wejść i wyjść, zmieniano liczbę warstw ukrytych, liczbę neuronów w warstwach ukrytych, liczbę epok uczenia, a także cztery parametry uczenia sieci: Learning rate, Tolerated error, Momentum i Flat spot elimination constant. Zmiany te miały na cele wyznaczenie optymalnego zestawu cech sieci neuronowej, która będzie dawała wyniki o jak najmniejszych różnicach pomiędzy wartościami przewidzianymi, a rzeczywistymi.

### *5.4.3. Wybrane wskaźniki oceny modelu.*

Do oceny poszczególnych wariantów korzystano ze wskaźników opisanych w podpunkcie 2.3.5. tj.: MSE, RMSE, MRE, $r^2$, $R^2$, wyliczanych przez program osobno dla zbioru uczącego i testowego. Przy porównywaniu poszczególnych wariantów modelu szczególną uwagę zwracano na MRE, w dalszej kolejności na pozostałe wskaźniki. Po przeprowadzeniu procesu uczenia i prognozowania, każdorazowo zapisywano wartości tych błędów. Dodatkowo, w ocenie modelu pomagał wykres, który powstawał, jako ostatni z elementów programu. Przedstawiał on rzeczywiste wartości danego wskaźnika (np. liczby wypadków) dla poszczególnych odcinków oraz wartości prognozowane przez sieć, z zaznaczeniem zbioru uczącego i testowego.

## *5.5. Testowanie poszczególnych wariantów modelu.*

Mając ustalone zarówno warstwy wejściowe dla sieci neuronowych, jak i ich warstwy wyjściowe, można było przystąpić do budowania poszczególnych modeli. Wszystkie one poddane zostały ocenie zgodnie z przyjętymi założeniami.

### *5.5.1. Proces testowania na przykładzie modelu z 28 zmiennymi wejściowymi.*

Na początku testowano modele, w których prognozowano liczby bezwzględne (5.2.I) na podstawie kompletu 28 zmiennych (5.3.3.A). Spośród kilkudziesięciu modelów, które przetestowano, kilka najlepszych przyniosło zbliżone wyniki (tabela w załączniku). Były to modele z 1 warstwą ukrytą, liczącą 15 neuronów. Znacznie gorsze rezultaty przynosiły sieci nazbyt rozbudowane. Z drugiej strony również sieci jednowarstwowe o niewielkiej liczbie neuronów nie osiągały równie dobrych wyników.

Optymalna liczba epok uczenia wyniosła 500. Przy zwiększaniu liczby epok, w poszczególnych przypadkach poprawie ulegały wskaźniki dla zbioru uczącego. W tym czasie również poprawiały się wyniki dla zbioru testowego, jednak tylko do określonego momentu. Po przekroczeniu optymalnej liczby epok, wyniki dla zbioru testowego zaczynały ulegać pogorszeniu.

Jeśli chodzi o pozostałe parametry, sieć neuronowa najlepsze efekty dawała przy wartości momentum równej 0. Zmiany pozostałych dostępnych parametrów przynosiły niewielki efekt. Należy też zauważyć, że ze względu na pewną losowość w procesie uczenia sieci neuronowych, powtarzanie prób przy braku zmiany parametrów mogły przynosić nieco różne wyniki. Stąd zapisywano najlepszy osiągnięty wynik.

Spośród 4 prognozowanych wskaźników: liczby wypadków, liczby rannych, ciężko rannych i zabitych, największą dokładnością cechowały się prognozy liczy wypadków. Niewiele mniej dokładne wyniki okazały się być dla liczby ofiar śmiertelnych i w dalszej kolejności liczby rannych. Zdecydowanie najgorzej sieć prognozowała liczbę ciężko rannych. Może to wskazywać na fakt, iż wielkość ta zmienia się w sposób najbardziej losowy – w niewielkim stopniu zależny od uwzględnionych czynników. W niektórych przypadkach, udawało się, zmieniając parametry sieci, polepszyć jeden ze wskaźników, kosztem innych. Nie były to jednak zmiany bardzo istotne.

Ostatecznie, na podstawie analizy wskaźników, za najdokładniejszy model prognozujący 4 wskaźniki dotyczące liczb bezwzględnych przy użyciu 28 zmiennych wejściowych, na raz należało uznać model o następujących cechach:

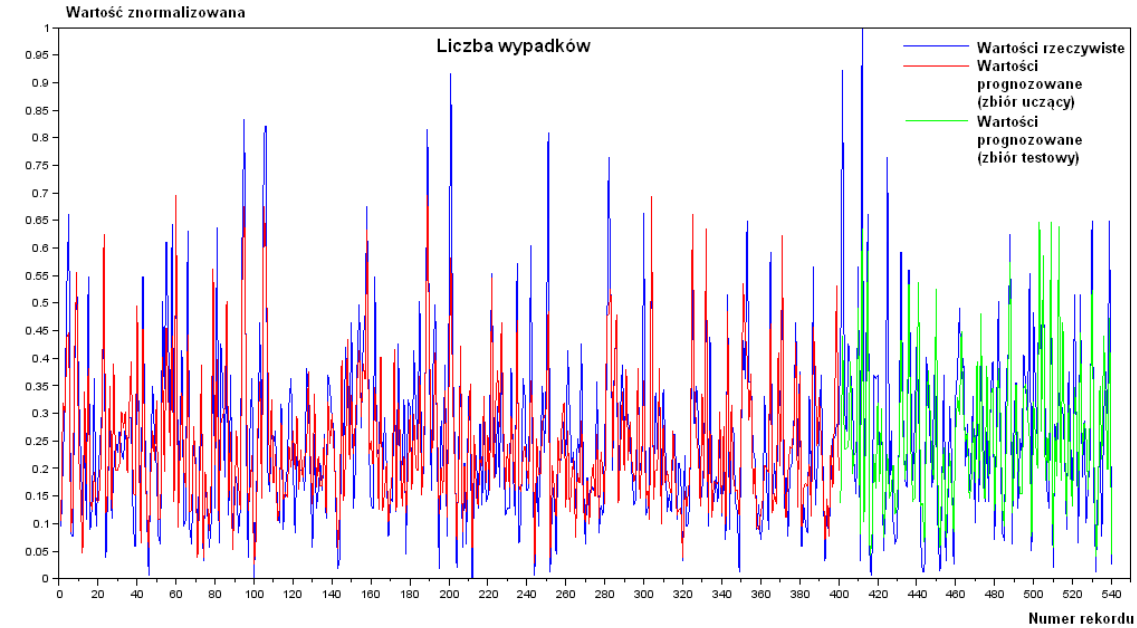
- 1 warstwa ukryta z 15 neuronami,
- 500 epok uczenia,
- pozostałe parametry, kolejno 0,1; 0.1; 0; 0.

Wyniki zbiorcze przedstawiono w tabeli 5.4. Zwraca uwagę fakt, że dla każdej zmiennej wyjściowej, wyniki na zbiorze testowym są gorsze od wyników na zbiorze uczącym. Chcąc zastosować w praktyce stworzony model, należy brać pod uwagę właśnie wyniki zbioru testowego, czyli tego w którym sieć nie znała wartości wyjściowych.

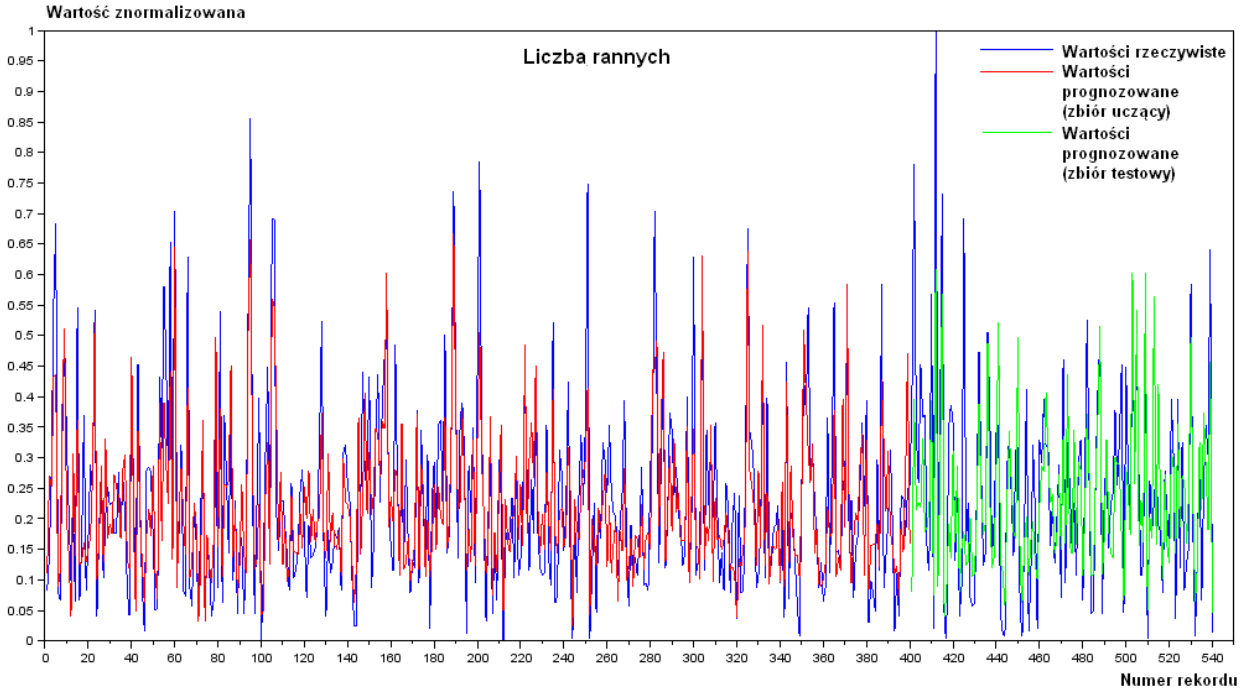
Tabela 5.3. Wyniki najlepszego modelu uwzględniającego wszystkie zmienne, przy 4 wyjściach w postaci liczb bezwzględnych

| | Zbiór uczący | | | | Zbiór testowy | | | |
|---|---|---|---|---|---|---|---|---|
| | **L. wypadków** | **L. rannych** | **L.c. rannych** | **L. zabitych** | **L. wyp.** | **L. rannych** | **L.c. rannych** | **L. zabitych** |
| **MSE** | 0,007694 | 0,00769 | 0,014332 | 0,0127096 | 0,0125984 | 0,0124552 | 0,0215522 | 0,0138243 |
| **RMSE** | 0,0877127 | 0,0876934 | 0,1197149 | 0,1142946 | 0,1122427 | 0,1116029 | 0,1468068 | 0,1175768 |
| **MRE** | 32,607 | 35,972706 | 44,448962 | 39,170375 | 36,301502 | 41,555795 | 60,175802 | 40,754566 |
| **$r^2$** | 0,907319 | 0,8854755 | 0,7253323 | 0,851533 | 0,8628603 | 0,8320058 | 0,6177044 | 0,8668681 |
| **$R^2$** | 0,7404636 | 0,6583418 | 0,3386149 | 0,5423008 | 0,6627704 | 0,5827207 | 0,3082566 | 0,6664991 |

Wyniki dla poszczególnych odcinków można odczytać z wykresów 5.4. – 5.7. Kolor niebieski oznacza rzeczywiste wartości, kolor czerwony wartości przewidziane przez sztuczną sieć neuronową (zbiór uczący), a zielony wartości dla zbioru testowego.

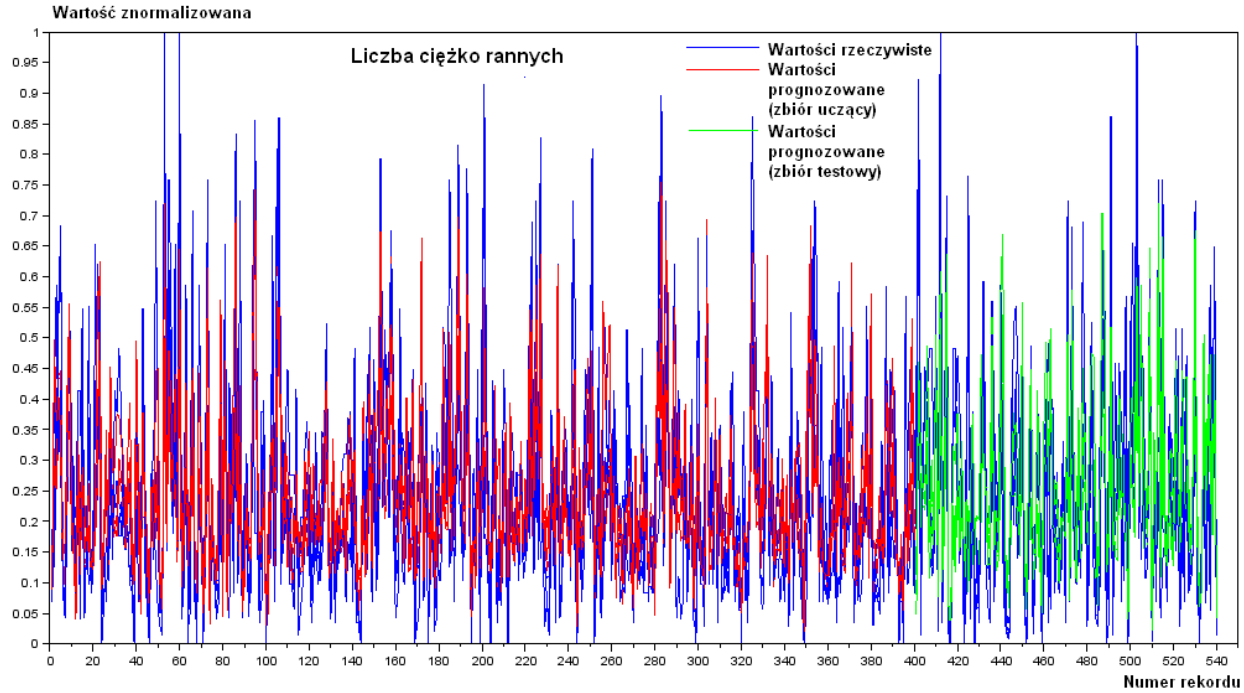



Rysunek 5.4. Wykres prognozy liczby wypadków w optymalnym modelu z 4 wyjściami i 28 wejściami



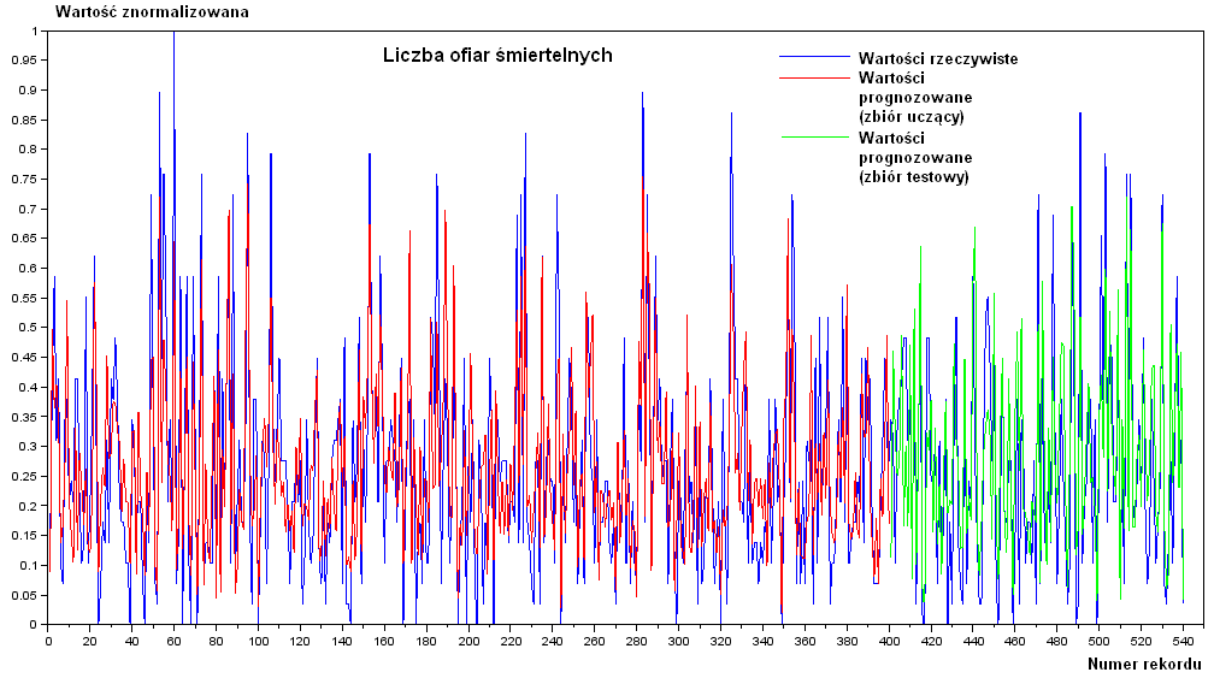

Rysunek 5.5. Wykres prognozy liczby rannych w optymalnym modelu z 4 wyjściami i 28 wejściami

Wykresy liczby wypadków i rannych mają zbliżony przebieg – na większości odcinków przyjmowały wartość znormalizowaną w okolicach 0,25. Niektóre odcinki cechuje znacznie większa liczba rannych i wypadków, brak natomiast odcinków, które w takim stopniu byłyby bezpieczniejsze od średnich.




Rysunek 5.6. Wykres prognozy liczby ciężko rannych w optymalnym modelu z 4 wyjściami i 28 wejściami

Powyższy wykres liczby ciężko rannych odróżnia się od pozostałych. Znacznie mniejsza jest w tym przypadku liczba obserwacji w okolicach średniej. Częściej natomiast przyjmowane są wartości w pobliżu maksymalnej. Może to być jedna z przyczyn gorszych wyników modelu dla tej cechy. Pod względem skupienia obserwacji, wykres liczby ofiar śmiertelnych można określić, jako pośredni między wykresem liczby ciężko rannych i pozostałymi.




Rysunek 5.7. Wykres prognozy liczby zabitych w optymalnym modelu z 4 wyjściami i 28 wejściami

Analizując wykresy, można zauważyć, że sieć dobrze radziła sobie z prognozowaniem wartości bliskich średniej. Mocno odstające wyniki były prognozowane zazwyczaj mniej dokładnie. Widać również wyraźną różnicę między dokładnością prognozy liczby ofiar ciężko rannych i prognozami pozostałych wskaźników. Ciężko natomiast dostrzec wizualnie różnice pomiędzy rezultatami w zbiorze uczącym, a testowym.

W dalszej kolejności testowano modele prognozujące osobno liczbę wypadków (5.2.IV), a następnie liczbę ofiar ciężko rannych (5.2.V) i zabitych (5.2.VI) na danym odcinku. Osiągnięto wyniki co najwyżej nieznacznie lepsze, niż w przypadku jednoczesnego prognozowania wszystkich cech. W związku z tym, można wysnuć wniosek, że przy chęci dokładniejszego przewidywania jednej tylko wartości, może okazać się opłacalne zbudowanie osobnego model. Porównanie z modelami prognozującymi jednocześnie wszystkie wskaźniki, zamieszczono w tabeli 5.5.

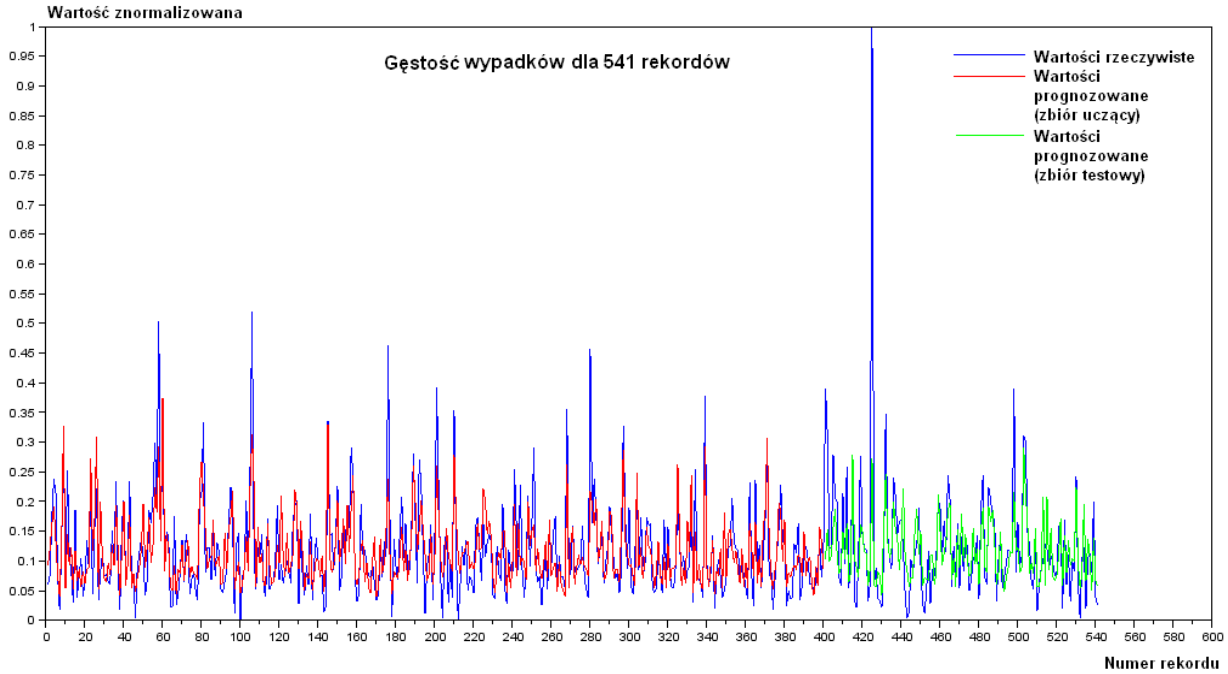
Tabela 5.4. Porównanie wyników najlepszych modeli uwzględniających wszystkie zmienne, przy 4 wyjściach w postaci liczb bezwzględnych i przy 1 wyjściu

| | Zbiór uczący | | | | Zbiór testowy | | | |
|---|---|---|---|---|---|---|---|---|
| | **Liczba wypadków** | | **Liczba zabitych** | | **Liczba wypadków** | | **Liczba zabitych** | |
| **MSE** | 0,007694 | 0,0080602 | 0,0127096 | 0,0134461 | 0,0125984 | 0,0127835 | 0,0138243 | 0,013846 |
| **RMSE** | 0,0877127 | 0,0897787 | 0,1142946 | 0,1159571 | 0,1122427 | 0,1130641 | 0,1175768 | 0,1142017 |
| **MRE** | 32,607 | 32,098512 | 39,170375 | 39,693833 | 36,301502 | 35,228537 | 40,754566 | 39,490894 |
| **$r^2$** | 0,907319 | 0,8929760 | 0,851533 | 0,8471159 | 0,8628603 | 0,8490288 | 0,8668681 | 0,8747955 |
| **$R^2$** | 0,7404636 | 0,6676446 | 0,5423008 | 0,568057 | 0,6627704 | 0,6083285 | 0,6664991 | 0,6773083 |
| | **Model z 4 wyjściami** | **Model z 1 wyjściem** | **Model z 4 wyjściami** | **Model z 1 wyjściem** | **Model z 4 wyjściami** | **Model z 1 wyjściem** | **Model z 4 wyjściami** | **Model z 1 wyjściem** |

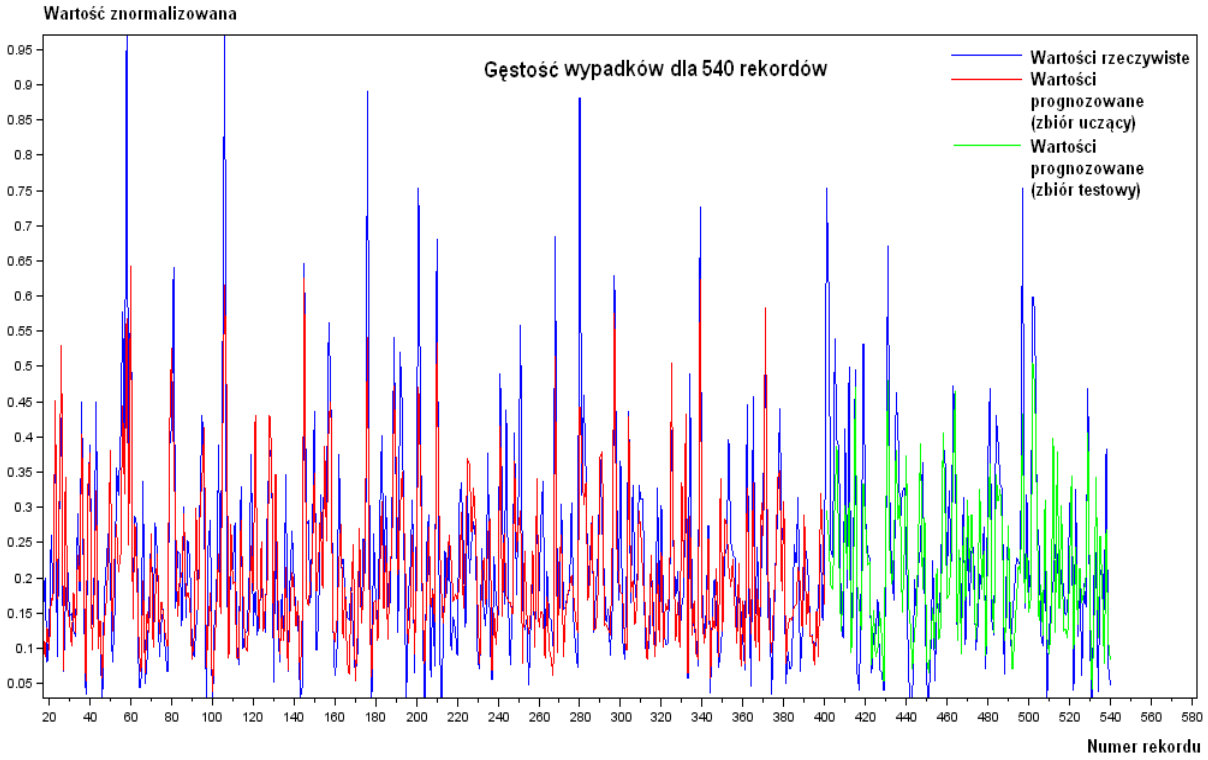
Najlepsze modele służące przewidywaniu jednej zmiennej wyraźnie różniły się pod względem struktury, w stosunku do poprzedniego optymalnego modelu. Przede wszystkim posiadały znacznie prostszą strukturę, składającą się z 1 warstwy liczącej 7 neuronów lub 10, dla modelu liczby zabitych. Również w tym przypadku, zwiększenie liczby warstw nie przyniosło znaczącej poprawy. Modele tego typu nie wymagały też większej liczby epok uczenia w celu osiągnięcia najlepszego wyniku. Wystarczająca okazywała się być liczba 300 epok.

Po modelach prognozujących liczby bezwzględne, opartych na wszystkich cechach wejściowych, testowano w analogiczny sposób modele gęstości i koncentracji wypadków i ofiar. W tym przypadku natrafiono na problem, jaki nie dotyczył poprzednich modeli. Jeden z odcinków (rekord 425 w zbiorze testowym - DK35-02) posiadał wartości zdecydowanie odstające od pozostałych – ponad 10 odchyleń standardowych od średniej. Gęstość wypadków była na nim niemal dwukrotnie większa, niż na drugim odcinku w tej klasyfikacji. Jego odmienność wskazała również analiza PCA. Zgodnie ze wskazaniami autorów podręcznika statystki [20] przeanalizowano możliwe przyczyny dużego odchylenia dla tego odcinka i podjęto decyzję o jego usunięciu ze zbioru danych. Odcinek ten posiadał duży udział terenu zabudowanego (ale nie odstający znacząco od innych) i powiązaną z nim dużą gęstość zjazdów. Stwierdzono jednak, że odstawanie wartości w tym przypadku mogło być także spowodowane przypadkową kumulacją zdarzeń w badanych latach na tym odcinku. W związku z tym, zdecydowano się uznać ten rekord za anormalny i postanowiono nie uwzględniać go w dalszych badań. Usunięcie odcinka poskutkowało radykalną poprawą dokładności prognoz, jeśli chodzi o współczynnik dopasowania i determinacji (pozostałe współczynniki okazały się mniej wrażliwe na takie wartości).

Na poniższych rysunkach 5.8. i 5.9. przedstawiono porównanie wyników dla modelu o tych samych cechach przy uwzględnieniu wszystkich 541 odcinków i po usunięcie wspomnianego odcinka.




Rysunek 5.8. Wyniki optymalnego modelu gęstości wypadków z 4 wyjściami i 28 wejściami przy zachowaniu 541 odcinków dróg




Rysunek 5.9. Wyniki optymalnego modelu gęstości z 4 wyjściami i 28 wejściami po usunięciu wyraźnie odstającego rekordu i pozostawieniu 540 odcinków dróg

W celu uniknięcia zafałszowania wyników, procedurę usuwania wyraźnie odstającego od pozostałych rekordu (i zmniejszania łącznej liczby prób do 540) kontynuowano również dla modeli gęstości o innej liczbie wejść i wyjść.

Porównując wyniki uzyskane dla modeli prognozujących liczby i gęstości zdarzeń przy komplecie zmiennych, można stwierdzić, że wiele zależy od przyjętej metody oceny. Biorąc pod uwagę $r^2$ i $R^2$, wyraźnie dokładniejsze od prognoz gęstości okażą się prognozy dotyczące wskaźników będących liczbami bezwzględnymi. Z kolei jeśli za istotniejsze uzna się MSE, RMSE, MRE, lepsze wyniki uzyskano dla gęstości wypadków, ofiar rannych, ciężko rannych. Wyjątkiem są prognozy gęstości ofiar zabitych.

Przyczyną różnic może być prawdopodobny istotny wpływ długości odcinka na wyniki. Zmienna ta traci tak istotne znaczenie przy modelowaniu wskaźników dotyczących gęstości zdarzeń. Z tego samego powodu, należy ocenić modele tego typu, jako bardziej praktyczne, pozwalające ocenić poziom bezpieczeństwa niezależnie od długości odcinka. Jeszcze nieco mniej dokładne okazują się prognozy koncentracji, jednak użycie tego wskaźnika ma kolejny atut w postaci uniezależnienia wyników nie tylko od długości odcinka, ale też od natężenia ruchu na nim. Można powiedzieć, że dopiero wykorzystanie wskaźników koncentracji pozwala dokładniej ocenić wpływ cech odcinka związanych z geometrią i organizacją, a nie wielkością ruchu i długością , zależną od przyjętego podziału.

### *5.5.2. Wyniki dla poszczególnych wariantów modelu.*

W kolejnych krokach, dokonano ograniczenia liczby zmiennych do wybranych 22 zmiennych (opisanych w podpunkcie 5.3.3.B), a następnie kolejno do 14 (C) i 3 (D). Dodatkowo testowano również model bez pracy przewozowej i natężenia ruchu, jako zmiennych objaśniających (E). W modelach tych zrezygnowano z osobnego prognozowania (przy pomocy jednego wyjścia) wskaźników dotyczących ciężko rannych, skupiając się na najważniejszych wskaźnikach bezpieczeństwa. W modelach o różnej liczbie wejść, optymalne rezultaty osiągano przy innej liczbie neuronów w warstwach ukrytych oraz liczbie epok uczenia. Kolejny raz potwierdza to fakt, iż do każdego modelu opartego o sztuczne sieci neuronowe należy podejść indywidualnie, zależnie od przyjętych cech wejściowych i wyjściowych.

Zestawienie obrazujące optymalną strukturę sieci neuronowej dla poszczególnych przypadków przedstawiono w poniższej tabeli 5.6. Kolorami zaznaczono modele najodpowiedniejsze dla poszczególnych wskaźników. Modele nr 23 i 104 przeprowadzają prognozę jednocześnie kilku wskaźników, jednak najlepiej przewidują odpowiednio: liczbę wypadków i ich koncentrację. Poza tym, jako model najlepiej prognozujący koncentrację zabitych, uznano przy komplecie wejść **wariant 103**. Zaznaczone modele zapisano w plikach załączonych do niniejszej pracy. Można z nich korzystać przy przeprowadzaniu dalszych badań przy pomocy środowiska Scilab i pakietu dotyczącego sztucznych sieci neuronowych, stosując własne zestawy wartości wejściowych. Dodatkowo, przedstawiono uproszczoną strukturę tych modeli na rys.5.10.

Tabela 5.5. Zestawienie cech optymalnych modeli o różnych wejściach i wyjściach sieci neuronowej

| L.p. | L. wyjść | Rodzaj wyjść | Wejścia | Liczba warstw ukrytych | Neurony w warstwach ukrytych | L. epok uczenia | Learning rate | Tolerated error | Momentum | Flat spot elimination constant |
|---|---|---|---|---|---|---|---|---|---|---|
| 23 | 4 | Liczby bezwzględne (wypadki) | wszystkie | 1 | 15 | 500 | 0,1 | 0,1 | 0 | 0 |
| 40 | 1 | Liczba wypadków | wszystkie | 1 | 7 | 300 | 0,1 | 0 | 0 | 0 |
| 45 | 1 | Liczba ciężko rannych | wszystkie | 1 | 7 | 300 | 0,1 | 0 | 0 | 0 |
| 48 | 1 | Liczba zabitych | wszystkie | 1 | 10 | 300 | 0,1 | 0 | 0 | 0 |
| 82 | 4 | Gęstości | wszystkie | 2 | 15 | 300 | 0,1 | 0,1 | 0 | 0 |
| 84 | 1 | Gęstość wypadków | wszystkie | 1 | 7 | 100 | 0,1 | 0,1 | 0 | 0 |
| 91 | 1 | Gęstość zabitych | wszystkie | 1 | 5 | 300 | 0,1 | 0,1 | 0 | 0 |
| 103 | 4 | Koncentracje (zabici) | wszystkie | 1 | 10 | 100 | 0,1 | 0 | 0 | 0 |
| 104 | 4 | Koncentracje (wypadki) | wszystkie | 1 | 10 | 100 | 0,1 | 0,2 | 0 | 0 |
| 110 | 1 | Koncentracja wypadków | wszystkie | 1 | 5 | 300 | 0,1 | 0,1 | 0 | 0 |
| 115 | 1 | Koncentracja zabitych | wszystkie | 2 | 5 | 300 | 0,1 | 0,1 | 0 | 0 |
| 130 | 4 | Liczby bezwzględne | wybór B | 1 | 7 | 300 | 0,1 | 0,2 | 0 | 0 |
| 139 | 1 | Liczba wypadków | wybór B | 1 | 10 | 500 | 0,1 | 0,1 | 0 | 0 |
| 153 | 1 | Liczba zabitych | wybór B | 1 | 5 | 500 | 0,1 | 0 | 0 | 0 |
| 171 | 4 | Gęstości | wybór B | 2 | 3 | 300 | 0,1 | 0,1 | 0 | 0 |
| 178 | 1 | Gęstość wypadków | wybór B | 1 | 7 | 500 | 0,1 | 0 | 0 | 0 |
| 184 | 1 | Gęstość zabitych | wybór B | 1 | 10 | 500 | 0,1 | 0 | 0 | 0 |
| 196 | 4 | Koncentracje | wybór B | 1 | 7 | 500 | 0,1 | 0 | 0 | 0 |
| 208 | 1 | Koncentracja wypadków | wybór B | 1 | 5 | 300 | 0,1 | 0 | 0 | 0 |
| 221 | 1 | Koncentracja zabitych | wybór B | 1 | 7 | 500 | 0,1 | 0 | 0 | 0 |
| 231 | 4 | Liczby bezwzględne | wybór C | 1 | 5 | 300 | 0,1 | 0 | 0 | 0 |
| 243 | 1 | Liczba wypadków | wybór C | 1 | 5 | 300 | 0,1 | 0 | 0 | 0 |
| 257 | 1 | Liczba zabitych | wybór C | 1 | 7 | 500 | 0,1 | 0 | 0 | 0 |
| 266 | 4 | Gęstości | wybór C | 1 | 7 | 1000 | 0,1 | 0 | 0 | 0 |
| 275 | 1 | Gęstość wypadków | wybór C | 1 | 5 | 300 | 0,1 | 0 | 0 | 0 |
| 283 | 1 | Gęstość zabitych | wybór C | 2 | 2 | 500 | 0,1 | 0 | 0 | 0 |
| 298 | 4 | Koncentracje | wybór C | 2 | 3 | 1000 | 0,1 | 0 | 0 | 0 |
| 302 | 1 | Koncentracja wypadków | wybór C | 1 | 5 | 500 | 0,1 | 0 | 0 | 0 |
| 314 | 1 | Koncentracja zabitych | wybór C | 1 | 7 | 500 | 0,1 | 0 | 0 | 0 |
| 323 | 4 | Liczby bezwzględne | wybór D | 1 | 3 | 1000 | 0,1 | 0 | 0 | 0 |

| L.p. | Liczba wyjść | Rodzaj wyjść | Wejścia | Liczba warstw ukrytych | Neurony w warstwach ukrytych | L. epok uczenia | Learning rate | Tolerated error | Momentum | Flat spot elimination constant |
|---|---|---|---|---|---|---|---|---|---|---|
| 334 | 1 | Liczba wypadków | wybór D | 1 | 5 | 500 | 0,1 | 0 | 0 | 0 |
| 340 | 1 | Liczba zabitych | wybór D | 1 | 10 | 500 | 0,1 | 0 | 0 | 0 |
| 350 | 4 | Gęstości | wybór D | 1 | 3 | 1000 | 0,1 | 0 | 0 | 0 |
| 360 | 1 | Gęstość wypadków | wybór D | 1 | 5 | 800 | 0,1 | 0 | 0 | 0 |
| 364 | 1 | Gęstość zabitych | wybór D | 1 | 5 | 800 | 0,1 | 0 | 0 | 0 |
| 371 | 4 | Koncentracje | wybór D | 1 | 7 | 800 | 0,1 | 0,2 | 0 | 0 |
| 378 | 1 | Koncentracja wypadków | wybór D | 1 | 5 | 300 | 0,1 | 0,2 | 0 | 0 |
| 392 | 1 | Koncentracja zabitych | wybór D | 1 | 3 | 1500 | 0,1 | 0 | 0 | 0 |
| 401 | 4 | Liczby bezwzględne | wybór E | 2 | 3 | 1000 | 0,1 | 0 | 0 | 0 |
| 407 | 4 | Gęstości | wybór E | 1 | 3 | 800 | 0,1 | 0,1 | 0 | 0 |
| 417 | 4 | Koncentracje | wybór E | 2 | 3 | 1000 | 0,1 | 0 | 0 | 0 |

Na podstawie tabeli można stąd wyciągnąć następujące wnioski:

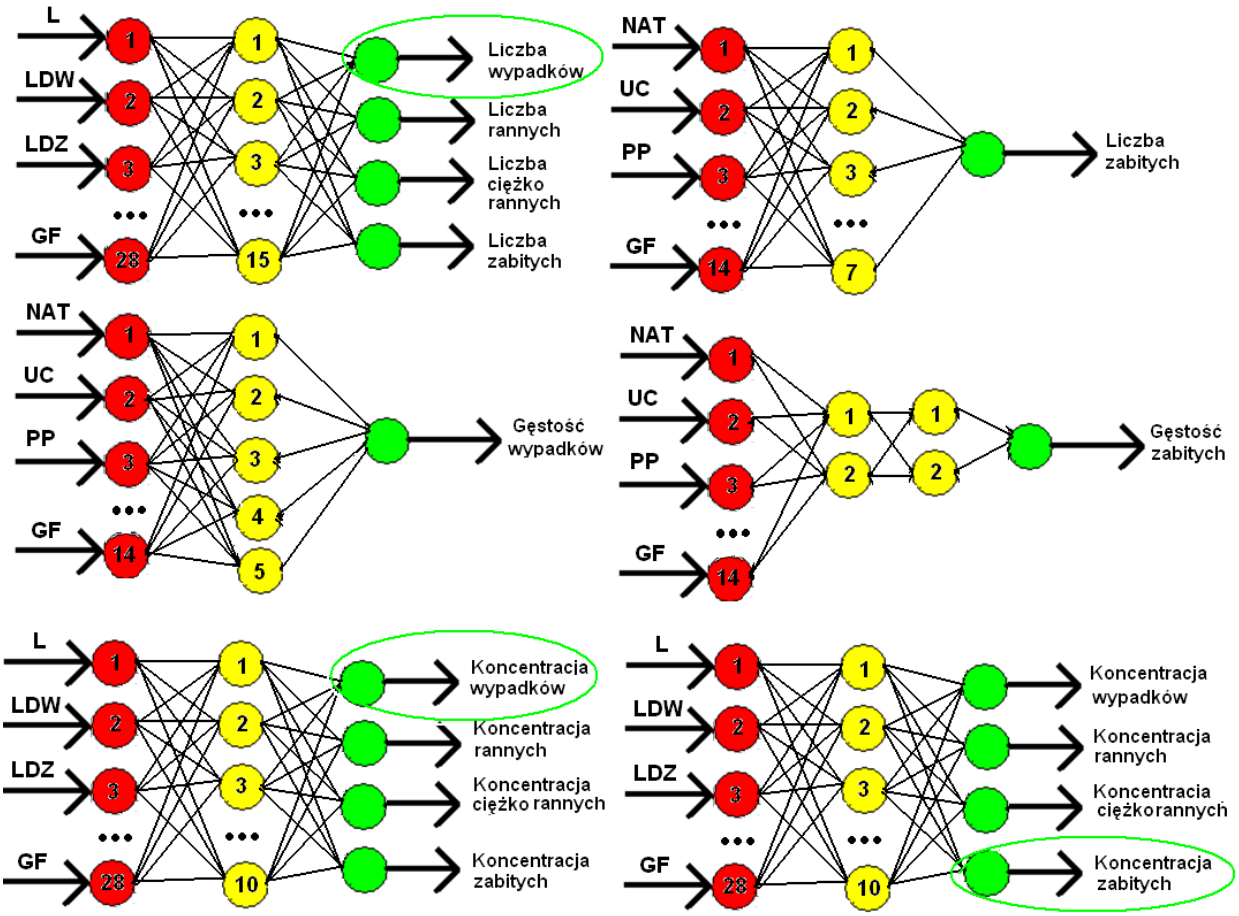
- w zdecydowanej większości przypadków optymalna okazywała się być struktura z jedną warstwą ukrytą. Kilka modeli najlepsze wyniki osiągało przy 2 warstwach ukrytych,

- zazwyczaj potwierdzało się, że im mniejsza liczba wejść sieci, tym mniejszej liczby neutronów w warstwach ukrytych potrzeba, by uzyskać najlepsze rezultaty,

- prawidłowość ta nie miała zastosowania, jeśli chodzi o zależność liczby neuronów na wyjściu i w warstwie ukrytej,

- sieci o prostej strukturze zasadniczo wymagały mniejszej liczby epok uczenia aby osiągnąć optymalne efekty, niż sieci o strukturze bardziej rozbudowanej,

- wartości momentum i flat spot elimination constant dla wszystkich najlepszych wariantów modelu były równe 0. Z kolei Tolerated error wahał się w granicach od 0 do 0,2.




Rysunek 5.10. Uproszczone schematy optymalnych modeli prognozujących poszczególne wskaźniki bezpieczeństwa.

.

Na powyższym rysunku na czerwono zaznaczono neurony wejściowe wraz ze skrótowym oznaczeniem cechy objaśniającej. Na żółto przedstawiono neurony w warstwie (lub warstwach ukrytych). Zielone punkty obrazują neurony wyjściowe ze zmienną prognozowaną, której dotyczą. W przypadku modeli optymalnych dla wybranego wskaźnika, ale prognozujących kilka wskaźników jednocześnie, zieloną obwódką zaznaczono optymalnie szacowany wskaźnik.

Dla poszczególnych modeli opisanych w tabeli 5.6. i przedstawionych na rys. 5.10., stworzono zestawienie wyników z uwzględnieniem różnych wskaźników oceny. Ograniczono się przy tym do wyników dla zbioru testowego. Zestawienie wyników zawarte jest w tabeli 5.7.

Tabela 5.6. Zestawienie wartości wskaźników oceny dla optymalnych modeli o różnych wejściach i wyjściach sieci neuronowej

| | **WEJŚCIA:** | Wszystkie | Wybór B | Wybór C | Wybór D | Wybór E |
|---|---|---|---|---|---|---|
| **Liczba wypadków** | MSE | 0,01260 | 0,01434 | 0,01338 | 0,01578 | 0,01313 |
| | MRE | 36,30 | 37,41 | 37,82 | 39,72 | 35,10 |
| | r² | 0,863 | 0,834 | 0,835 | 0,809 | 0,782 |
| | R² | 0,663 | 0,484 | 0,503 | 0,487 | 0,246 |
| **Liczba zabitych** | MSE | 0,01385 | 0,01357 | 0,01379 | 0,01502 | 0,02479 |
| | MRE | 39,86 | 38,43 | 38,04 | 40,18 | 41,55 |
| | r² | 0,874 | 0,867 | 0,866 | 0,850 | 0,751 |
| | R² | 0,712 | 0,631 | 0,638 | 0,572 | 0,148 |
| **Gęstość wypadków** | MSE | 0,00954 | 0,00971 | 0,00859 | 0,01603 | 0,22647 |
| | MRE | 33,47 | 39,85 | 35,29 | 40,25 | 46,40 |
| | r² | 0,846 | 0,824 | 0,850 | 0,807 | 0,657 |
| | R² | 0,470 | 0,500 | 0,502 | 0,498 | 0,189 |
| **Gęstość zabitych** | MSE | 0,01386 | 0,00607 | 0,01214 | 0,01454 | 0,01976 |
| | MRE | 40,76 | 45,85 | 39,23 | 40,60 | 59,80 |
| | r² | 0,818 | 0,783 | 0,837 | 0,850 | 0,409 |
| | R² | 0,459 | 0,511 | 0,487 | 0,578 | 0,100 |
| **Koncentracja wypadków** | MSE | 0,00978 | 0,01282 | 0,01329 | 0,01777 | 0,02430 |
| | MRE | 30,16 | 34,99 | 34,73 | 35,47 | 34,50 |
| | r² | 0,858 | 0,801 | 0,779 | 0,796 | 0,794 |
| | R² | 0,423 | 0,359 | 0,243 | 0,178 | 0,223 |
| **Koncentracja zabitych** | MSE | 0,02336 | 0,02574 | 0,02422 | 0,16406 | 0,02587 |
| | MRE | 38,49 | 40,74 | 39,06 | 39,61 | 39,41 |
| | r² | 0,774 | 0,763 | 0,763 | 0,770 | 0,759 |
| | R² | 0,134 | 0,283 | 0,157 | 0,165 | 0,112 |

Z analizy tabeli wyników, można wyciągnąć następujące wnioski:

- ocena poszczególnych wyniki w bardzo dużym stopniu zależy od przyjętych wskaźników oceny. W wielu przypadkach MSE, MRE, $r^2$ i $R^2$ wskazują na inny model, jako najlepszy,

- przy wszystkich wejściach, najdokładniejszą prognozę udało się uzyskać dla gęstości wypadków. Wskazują na to wyniki wszystkich wskaźników, poza $R^2$, dla którego najlepiej wypadła prognoza liczby zabitych,

- przy wyborze B, pod kątem MSE najdokładniejsza była prognoza gęstości zabitych, pod kątem MRE – prognoza koncentracji wypadków, a pod kątem $r^2$ i $R^2$ prognoza liczby zabitych,

- bardzo podobnie rozkład wyglądał przy wyborze C i D,

- przy wyborze E, zależnie od wskaźnika, za najlepsze należy uznać prognozy liczby wypadków lub koncentracji wypadków,

- określenie najlepszego wyboru zestawu zmiennych również zależy od wybrania wskaźnika oceny i zmiennej celu. Zasadniczo jednak najlepszy okazywał się wybór C lub kompletny zestaw wejść.

### *5.5.3. Próba poprawy dokładności wybranego wariantu modelu.*

W poprzednim podpunkcie przedstawiono wyniki dla najlepszych wariantów modelu. Posiadają one określoną optymalną strukturę i przyjętą z góry liczbę epok nauczania. W celu poprawy osiąganych przez nie rezultatów, można dokonać wykonać dodatkowe czynności:

- ponowne przeprowadzenie procesu uczenia z monitorowaniem błędu i w efekcie dokładniejsze wyznaczenie optymalnej liczby epok uczenia,

- zmniejszenie liczby przypadków (odcinków dróg) określonych zestawem wartości cech podawanych na wejściu sieci neuronowej [15].

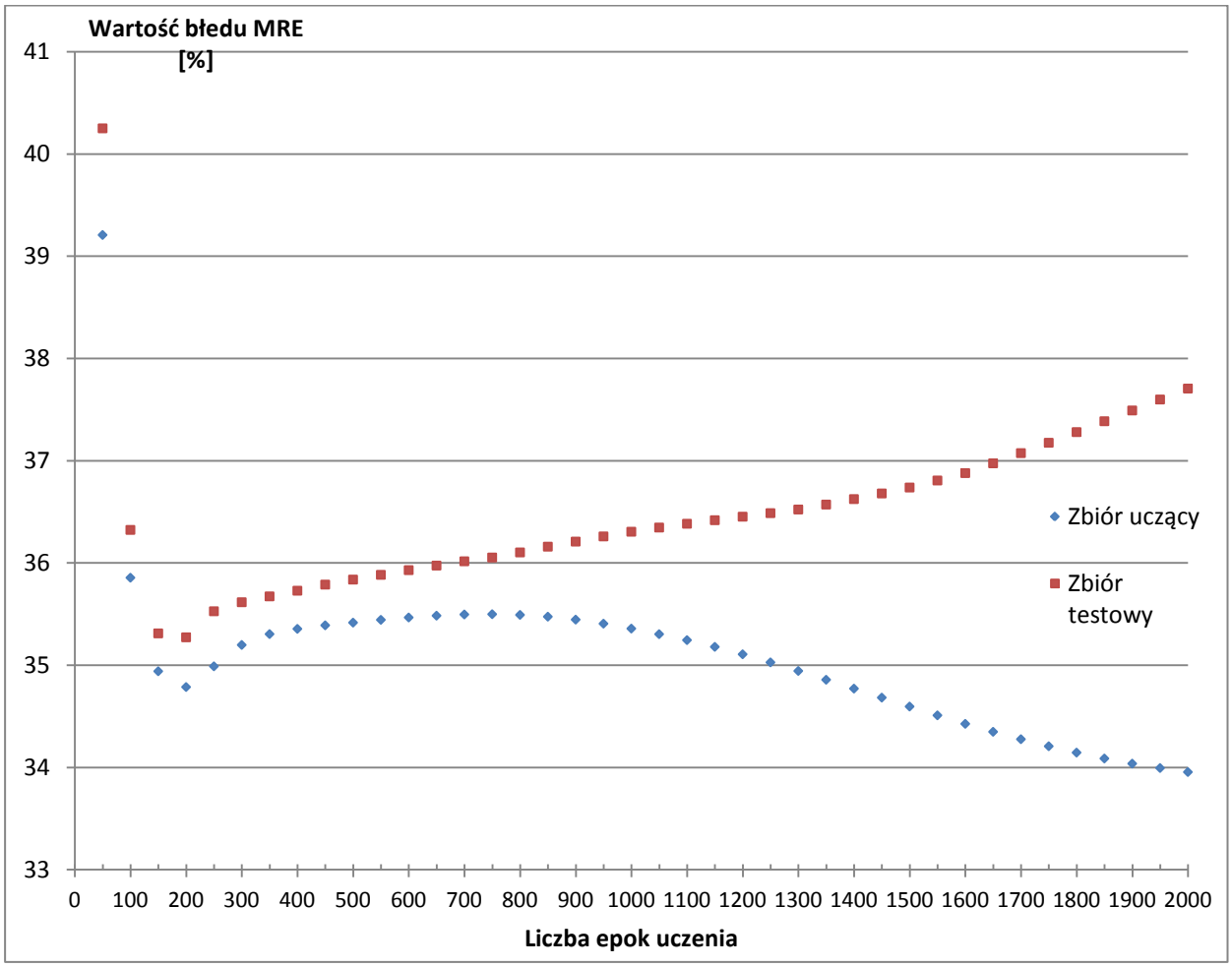
Czynności te, ograniczono do wybranego wariantu modelu. Będzie to wariant nr 275, opisany w tabeli 5.6. Posiada on jedne z niższych wartości MRE i MSE spośród wszystkich modeli (przedstawione w tabeli 5.7.) oraz stosunkowo wysoki współczynnik $r^2$. Ponadto gęstość wypadków, która jest w tym przypadku prognozowana jest popularnym wskaźnikiem eliminującym wpływ długości odcinka na wyniki.

Monitorowanie procesu uzyskano dzięki modyfikacji używanego wcześniej algorytmu w środowisku Scilab. W tym przypadku przeprowadzono prognozowanie w oparciu o naukę z liczbą epok od 0 do 2000 z krokiem równym 50 epok. Przy tym po każdym kroku, program zachowywał wartości wskaźników błędów zarówno dla zbioru uczącego, jak i testowego. Co istotne, po każdym kroku następowała kontynuacja uczenia – kolejne 50 epok, a nie rozpoczynanie procesu od nowa.

Wnioski na temat otrzymanych rezultatów najłatwiej wyciągnąć na podstawie rysunku. Na rys. 5.10. przedstawiono przykładowy wykres, zawierający wartość błędu MRE zbioru uczącego i testowego dla kolejnych epok uczenia. Nie trudno stwierdzić, że wstępne stwierdzenia wysunięte na początku punktu 5.5.1. były prawidłowe.

Wraz z dodawaniem kolejnych epok uczenia, błąd dla zbioru uczącego spada w sposób dość regularny. Tylko pomiędzy 200 a 700 epoką zanotowano minimalny,

symboliczny wzrost błędu MRE dla zbioru uczącego. Z kolei błąd na zbiorze testowym początkowo spada, aż do optymalnej liczby epok uczenia. Następnie przy wydłużeniu procesu uczenia błąd zaczyna wyraźnie rosnąć, w przybliżeniu w sposób liniowy. W każdym punkcie wykresu, błąd na zbiorze testowym jest większy od błędu na zbiorze uczącym, jednak różnica ta powiększa się wyraźnie po przekroczeniu optymalnej liczby epok.




Rysunek 5.10. Wyniki dla zbioru uczącego i testowego wybranego modelu w zależności od liczby epok uczenia

Korzystając z wykresu można również stwierdzić, że optymalna liczba epok uczenia ustalona wcześniej na 300 wynosi nieco mniej – 200 epok (przy założonej częstotliwości próbkowania 50). Oznacza to, że udało się w pewien sposób polepszyć uzyskane wcześniej wyniki.

Innym sposobem na poprawę dokładności prognoz sztucznej sieci neuronowej może być zmiana liczby rekordów wykorzystanych do uczenia. We wszystkich wcześniejszych modelach zbiór uczący zawierał 400 odcinków, a testowy 141 lub 140 odcinków. Dla wspomnianej powyżej struktury (wariant 285) sprawdzono działanie modelu nie tylko przy 400, ale też 300, 200 i 100 dobranych losowo odcinkach w zbiorze uczącym. Zbiór testowy pozostał bez zmian.

W tabeli 5.8. zestawiono otrzymane wyniki. Widać z nich, że sieć najlepiej funkcjonuje wcale nie przy największej liczbie danych. Wszystkie wskaźniki osiągały wartości najkorzystniejsze dla 300 elementów zbioru uczącego. Przy 100 odcinkach zanotowano wyniki najgorsze, co może oznaczać, że liczba danych była niewystarczająca, aby dostatecznie nauczyć sieć neuronową. Z drugiej strony, maksymalna dostępna liczba danych uczących (400), okazała się mniej korzystna. Przyczyną może być „uczenie na pamięć" i utrata zdolności do uogólniania wyników przez sztuczną sieć neuronową [15].

Tabela 5.7. Wyniki dla zbioru testowego modelu gęstości zabitych przy różnej liczbie elementów zbioru uczącego

| | **Liczba danych** | **400** | **300** | **200** | **100** |
|---|---|---|---|---|---|
| **Gęstość wypadków** | MSE | 0,00860 | 0,00842 | 0,01171 | 0,01293 |
| | RMSE | 0,09272 | 0,09174 | 0,10820 | 0,11372 |
| | MRE | 35,27 | 33,09 | 36,22 | 38,26 |
| | $r^2$ | 0,849 | 0,875 | 0,825 | 0,789 |
| | $R^2$ | 0,473 | 0,503 | 0,395 | 0,250 |

Ostatecznie można stwierdzić, że również w tym wypadku udało się poprawić osiągane wyniki. Należy przy tym zaznaczyć, iż optymalna liczba danych może być różna dla poszczególnych wariantów modelu i każdy przypadek należałoby rozważyć oddzielnie.

# 6. OCENA I ANALIZA OTRZYMANYCH REZULTATÓW.

Kwestią kluczową z punktu widzenia oceny zasadności zastosowania sztucznych sieci neuronowych do prognoz bezpieczeństwa na odcinkach dróg jest interpretacja uzyskanych wyników. Dokonano jej w oparciu o analizę uzyskanych wartości miar dokładności oraz porównaniu ich z innymi modelami. Skupiono się przy tym oczywiście tylko na tych modelach, które okazały się najlepsze do prognoz poszczególnych wskaźników. Strukturę tych modeli przedstawiono w tabeli 5.6., a wyniki w tabeli 5.7.

Poza prognozowaniem poziomu bezpieczeństwa, innym ważnym celem pracy było określenie wpływu poszczególnych cech drogi na powstawaniu wypadków. W oparciu o stworzone modele, dokonano analizy pozwalającej określić, jak istotna jest dana cecha z punktu widzenia bezpieczeństwa.

## *6.1. Analiza otrzymanych wyników i porównanie z innymi modelami.*

Otrzymane wyniki można analizować poprzez ich porównanie z rzeczywistymi wartościami. Takie zestawienie przedstawiono w tabeli 6.1.

Powszechnie używanym wskaźnikiem oceny dokładności prognoz w przypadku sieci neuronowych jest średni błąd względny [8,14]. Wartość tego błędu przy dość dobrej prognozie wynieść może ok. 30% [14]. Jeden z najmniejszych uzyskanych w stworzonych modelach błąd MRE wyniósł 33,08% - dla prognoz gęstości po dobraniu odpowiedniej liczby danych i dokładnym określeniu optymalnej liczby epok uczenia. Dla modelu prognozującego koncentrację wypadków, wartość ta wyniosła nawet 30,16%, jednak przy znacznie gorszych wartościach pozostałych wskaźników. W innych modelach wartość MRE wahała się między 35, a 40%. Oznacza to, że dokładność prognoz jest zadowalająca. Przykłady innych badań pokazują jednak, że możliwe jest uzyskanie nawet MRE = 20% przy modelowaniu prognozowaniu poziomu bezpieczeństwa na drogach z wykorzystaniem sieci neuronowych.

Inna będzie ocena otrzymanych wyników, jeśli pod uwagę weźmie się wskaźnik $r^2$ lub $R^2$. Wartość tego pierwszego dla dokładnego modelu powinna wynosić blisko 1 [8]. Tymczasem najwyższą wartość, jaką udało się osiągnąć, to 0,866 dla prognozy liczby zabitych. Wartości $R^2$ są bardziej zróżnicowanie, jednak modele, w których osiągnięto $R^2$ ok. 0,6 (liczba wypadków, liczba zbitych) można ocenić, jako zadowalające [19]. Pozostałe wg tego wskaźnika są słabe (np. 0,487 dla gęstości zabitych) lub bardzo słabe (0,134 dla koncentracji zabitych).

Tabela 6.1. Porównanie prognoz modelu 23. i rzeczywistych wartości wskaźników bezpieczeństwa dla poszczególnych odcinków dróg

| | Liczba wypadków | | Liczba rannych | | liczba ciężkorannyc | | Liczba zabitych | | | Liczba wypadków | | Liczba rannych | | liczba ciężkorannyc | | Liczba zabitych | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Symbol odcinka | Rzeczywista | Prognozowana | Rzeczywista | Prognozowana | Rzeczywista | Prognozowana | Rzeczywista | Prognozowana | Symbol odcinka | Rzeczywista | Prognozowana | Rzeczywista | Prognozowana | Rzeczywista | Prognozowana | Rzeczywista | Prognozowana |
| DK65-01a | 2 | 11 | 2 | 19 | 0 | 6 | 0 | 1 | DK61-05 | 45 | 29 | 61 | 38 | 20 | 14 | 13 | 9 |
| DK42-03 | 3 | 21 | 4 | 25 | 2 | 9 | 1 | 2 | DK19-05 | 45 | 42 | 65 | 56 | 12 | 16 | 7 | 10 |
| DK68-01 | 3 | 9 | 3 | 12 | 1 | 6 | 1 | 5 | DK08-02 | 45 | 49 | 74 | 63 | 34 | 20 | 5 | 8 |
| DK22-08a | 3 | 8 | 3 | 12 | 1 | 8 | 0 | 2 | DK16-06 | 46 | 44 | 80 | 66 | 34 | 22 | 7 | 10 |
| S22-02 | 3 | 10 | 3 | 19 | 0 | 7 | 1 | 2 | DK11-06 | 47 | 33 | 63 | 43 | 12 | 15 | 4 | 7 |
| DK25-09a | 4 | 9 | 6 | 14 | 1 | 6 | 0 | 2 | DK08-11 | 47 | 35 | 66 | 49 | 9 | 16 | 6 | 9 |
| DK12-17 | 4 | 9 | 2 | 14 | 1 | 5 | 3 | 1 | DK03-09 | 47 | 45 | 82 | 61 | 28 | 18 | 3 | 6 |
| DK78-04 | 5 | 17 | 6 | 27 | 0 | 8 | 0 | 2 | DK28-07 | 47 | 68 | 65 | 92 | 24 | 22 | 10 | 9 |
| DK60-03a | 5 | 8 | 4 | 13 | 1 | 4 | 1 | 1 | DK39-02 | 48 | 32 | 60 | 39 | 18 | 15 | 1 | 4 |
| DK16-01 | 6 | 10 | 6 | 14 | 2 | 7 | 3 | 2 | DK22-01 | 48 | 30 | 75 | 45 | 35 | 13 | 8 | 6 |
| S01-02 | 6 | 11 | 5 | 20 | 2 | 9 | 5 | 4 | DK19-13 | 49 | 50 | 66 | 72 | 19 | 17 | 9 | 9 |
| DK16-02 | 7 | 12 | 13 | 17 | 2 | 7 | 0 | 2 | DK65-03 | 49 | 45 | 68 | 63 | 15 | 20 | 4 | 5 |
| DK25-01 | 9 | 12 | 11 | 17 | 0 | 7 | 2 | 3 | DK02-15 | 49 | 78 | 47 | 104 | 10 | 27 | 22 | 18 |
| DK10-01 | 9 | 14 | 12 | 16 | 2 | 7 | 0 | 1 | DK11-11 | 50 | 43 | 84 | 60 | 1 | 17 | 3 | 9 |
| DK60-05 | 11 | 20 | 15 | 26 | 9 | 11 | 5 | 6 | DK06-02 | 50 | 45 | 79 | 63 | 28 | 20 | 8 | 13 |
| DK15-04 | 11 | 17 | 12 | 25 | 0 | 8 | 0 | 3 | DK08-01 | 51 | 86 | 78 | 115 | 22 | 29 | 13 | 15 |
| DK62-01 | 11 | 25 | 10 | 36 | 2 | 12 | 2 | 5 | DK72-01 | 51 | 26 | 70 | 38 | 3 | 10 | 7 | 5 |
| DK20-05 | 12 | 20 | 13 | 27 | 5 | 10 | 2 | 4 | DK07-02 | 52 | 84 | 87 | 120 | 11 | 29 | 8 | 17 |
| DK63-05 | 12 | 16 | 17 | 23 | 2 | 9 | 2 | 4 | DK08-10 | 53 | 42 | 81 | 51 | 7 | 18 | 4 | 9 |
| DK78-03 | 12 | 19 | 13 | 28 | 2 | 9 | 1 | 2 | DK28-04 | 53 | 53 | 71 | 73 | 23 | 17 | 8 | 5 |
| DK66-03 | 13 | 18 | 16 | 26 | 3 | 9 | 1 | 4 | DK11-10 | 54 | 34 | 91 | 50 | 22 | 11 | 14 | 6 |
| DK14-05 | 13 | 24 | 12 | 31 | 2 | 11 | 4 | 4 | DK53-03 | 54 | 60 | 81 | 85 | 14 | 20 | 15 | 10 |
| DK22-07 | 13 | 35 | 17 | 48 | 9 | 15 | 3 | 6 | DK06-01 | 54 | 61 | 71 | 89 | 31 | 23 | 20 | 14 |
| DK19-01 | 14 | 18 | 12 | 20 | 3 | 11 | 4 | 5 | DK06-06 | 54 | 48 | 69 | 67 | 24 | 21 | 19 | 11 |
| DK01-01 | 15 | 17 | 20 | 25 | 5 | 12 | 5 | 6 | DK07-18 | 54 | 103 | 57 | 127 | 19 | 37 | 22 | 19 |
| DK39-03 | 17 | 34 | 18 | 47 | 6 | 14 | 0 | 5 | DK91-11 | 55 | 59 | 63 | 74 | 5 | 20 | 11 | 6 |
| DK57-07 | 17 | 24 | 18 | 30 | 6 | 13 | 1 | 3 | DK35-03b | 56 | 46 | 70 | 62 | 27 | 18 | 5 | 6 |
| DK58-04 | 17 | 45 | 21 | 61 | 8 | 16 | 3 | 5 | DK92-06a | 57 | 58 | 74 | 81 | 22 | 25 | 25 | 15 |
| DK63-12 | 19 | 19 | 44 | 25 | 6 | 9 | 1 | 3 | DK17-02 | 58 | 25 | 95 | 38 | 25 | 13 | 14 | 9 |
| DK61-04 | 21 | 26 | 23 | 37 | 6 | 13 | 9 | 6 | DK75-03 | 58 | 45 | 93 | 62 | 23 | 15 | 2 | 4 |
| DK74-11 | 21 | 27 | 38 | 36 | 1 | 12 | 3 | 3 | DK09-05 | 58 | 73 | 78 | 100 | 4 | 24 | 10 | 13 |
| DK32-06 | 21 | 28 | 33 | 41 | 12 | 14 | 6 | 6 | DK22-11 | 58 | 47 | 69 | 67 | 9 | 17 | 3 | 8 |
| DK57-08 | 22 | 25 | 27 | 38 | 13 | 10 | 5 | 3 | DK10-09 | 59 | 41 | 86 | 59 | 28 | 15 | 14 | 9 |
| DK40-01 | 23 | 27 | 24 | 36 | 6 | 10 | 2 | 2 | DK16-05 | 59 | 57 | 88 | 79 | 34 | 22 | 7 | 9 |
| DK58-03 | 23 | 28 | 39 | 35 | 7 | 11 | 4 | 3 | DK59-02 | 59 | 32 | 77 | 38 | 23 | 12 | 8 | 4 |
| DK12-04 | 23 | 32 | 27 | 37 | 9 | 13 | 4 | 4 | DK03-06 | 59 | 33 | 101 | 47 | 35 | 16 | 13 | 12 |
| DK22-03 | 23 | 20 | 39 | 29 | 22 | 11 | 6 | 4 | DK17-01 | 60 | 55 | 90 | 81 | 15 | 23 | 12 | 14 |
| DK45-02 | 25 | 37 | 29 | 50 | 10 | 14 | 9 | 5 | DK45-05 | 60 | 49 | 86 | 67 | 29 | 17 | 10 | 8 |
| DK16-09 | 25 | 36 | 45 | 51 | 12 | 14 | 2 | 4 | DK50-08 | 60 | 58 | 72 | 73 | 23 | 21 | 11 | 14 |
| DK66-01 | 26 | 22 | 41 | 29 | 3 | 11 | 5 | 6 | DK16-10 | 60 | 48 | 69 | 63 | 18 | 18 | 7 | 7 |
| DK45-04 | 26 | 23 | 32 | 35 | 6 | 10 | 5 | 3 | DK08-09 | 60 | 58 | 87 | 74 | 4 | 22 | 17 | 14 |
| DK32-05 | 26 | 38 | 27 | 52 | 13 | 17 | 6 | 6 | DK07-05 | 63 | 76 | 114 | 102 | 30 | 27 | 21 | 17 |
| DK45-01 | 27 | 38 | 36 | 48 | 7 | 11 | 1 | 3 | DK19-15 | 63 | 58 | 81 | 75 | 23 | 21 | 9 | 7 |
| DK02-04 | 27 | 33 | 35 | 44 | 10 | 17 | 13 | 11 | DK62-09 | 64 | 47 | 71 | 56 | 23 | 20 | 9 | 7 |
| DK05-03 | 27 | 32 | 45 | 46 | 13 | 16 | 6 | 9 | DK71-02 | 64 | 50 | 79 | 68 | 23 | 16 | 8 | 4 |
| DK43-01 | 28 | 27 | 32 | 37 | 7 | 10 | 5 | 4 | DK10-10 | 65 | 77 | 80 | 113 | 49 | 28 | 12 | 17 |
| DK60-01 | 28 | 20 | 34 | 29 | 13 | 8 | 4 | 2 | DK08-17 | 65 | 60 | 101 | 85 | 34 | 27 | 19 | 18 |
| DK82-02 | 29 | 27 | 33 | 39 | 28 | 13 | 7 | 4 | DK74-07 | 65 | 56 | 93 | 77 | 33 | 21 | 11 | 12 |
| DK51-01 | 29 | 28 | 31 | 38 | 11 | 11 | 2 | 3 | DK74-04 | 67 | 19 | 96 | 22 | 21 | 6 | 10 | 4 |
| DK01-03 | 29 | 22 | 34 | 30 | 21 | 14 | 7 | 7 | DK74-02 | 67 | 53 | 84 | 66 | 28 | 23 | 17 | 16 |
| DK08-22 | 30 | 38 | 30 | 53 | 5 | 19 | 11 | 13 | DK46-04 | 68 | 35 | 112 | 47 | 29 | 18 | 11 | 8 |
| DK19-04 | 32 | 34 | 35 | 47 | 8 | 15 | 14 | 8 | DK25-06 | 77 | 56 | 110 | 82 | 38 | 23 | 9 | 11 |
| DK14-02 | 33 | 37 | 39 | 53 | 6 | 13 | 3 | 5 | DK16-04 | 78 | 50 | 97 | 66 | 25 | 20 | 1 | 6 |
| DK79-05 | 35 | 27 | 52 | 34 | 24 | 13 | 8 | 6 | DK05-06 | 80 | 61 | 130 | 90 | 36 | 23 | 11 | 14 |
| DK25-04 | 35 | 24 | 48 | 36 | 12 | 11 | 3 | 4 | DK22-10 | 82 | 59 | 97 | 78 | 11 | 24 | 14 | 12 |
| DK46-09 | 36 | 35 | 44 | 40 | 12 | 12 | 6 | 4 | DK91-12 | 82 | 75 | 97 | 95 | 21 | 24 | 4 | 7 |
| DK12-15 | 36 | 49 | 38 | 60 | 11 | 20 | 7 | 8 | DK91-01 | 87 | 92 | 121 | 126 | 10 | 32 | 14 | 13 |
| DK03-02 | 37 | 24 | 60 | 35 | 20 | 15 | 7 | 10 | DK07-16b | 88 | 55 | 112 | 55 | 41 | 22 | 10 | 6 |
| DK05-14 | 37 | 33 | 44 | 41 | 9 | 15 | 2 | 4 | DK44-02 | 89 | 84 | 125 | 118 | 27 | 28 | 12 | 10 |
| DK10-11 | 39 | 36 | 50 | 49 | 19 | 17 | 10 | 10 | DK79-12 | 90 | 76 | 140 | 96 | 33 | 24 | 11 | 9 |
| DK02-01 | 40 | 52 | 59 | 66 | 25 | 22 | 10 | 15 | DK47-02 | 94 | 74 | 117 | 102 | 33 | 27 | 15 | 6 |
| DK07-07 | 40 | 44 | 59 | 56 | 32 | 20 | 16 | 8 | DK28-06 | 94 | 102 | 130 | 144 | 43 | 30 | 14 | 14 |
| DK11-07 | 40 | 52 | 65 | 74 | 9 | 22 | 7 | 14 | DK12-16 | 99 | 89 | 123 | 129 | 22 | 27 | 13 | 17 |
| DK03-08 | 40 | 56 | 80 | 77 | 20 | 25 | 8 | 13 | DK12-19 | 100 | 98 | 128 | 136 | 72 | 36 | 23 | 19 |
| DK31-04 | 41 | 23 | 56 | 31 | 33 | 12 | 6 | 5 | DK14-01 | 103 | 97 | 144 | 128 | 27 | 30 | 21 | 20 |
| DK06-05 | 41 | 30 | 51 | 42 | 25 | 16 | 10 | 8 | DK73-04 | 103 | 80 | 158 | 109 | 25 | 25 | 9 | 12 |
| DK63-10 | 43 | 40 | 64 | 55 | 21 | 18 | 10 | 6 | DK04-17 | 105 | 101 | 180 | 137 | 25 | 33 | 13 | 19 |
| DK02-20 | 43 | 25 | 56 | 34 | 21 | 15 | 11 | 10 | DK35-02 | 121 | 37 | 170 | 38 | 50 | 15 | 3 | 2 |
| DK79-08 | 43 | 61 | 56 | 80 | 11 | 20 | 7 | 10 | DK17-07 | 146 | 75 | 192 | 102 | 52 | 27 | 9 | 13 |
| DK62-08 | 45 | 21 | 50 | 29 | 13 | 12 | 6 | 4 | DK75-02 | 158 | 102 | 245 | 143 | 39 | 31 | 14 | 13 |
| DK16-08 | 45 | 46 | 56 | 53 | 12 | 16 | 5 | 4 | | | | | | | | | |

Kluczowe w ocenie stworzonych modelów jest porównanie ich wyników z modelem regresyjnym, opartym na tych samych danych i również dotyczącym dróg krajowych. Wskaźnik $R^2$ wyniósł tam 0,47 dla liczby zabitych na drogach jednojezdniowych i 0,44 dla gęstości zabitych na tych drogach [6]. Pozostałych wskaźników nie uwzględniono. W związku z tym, można jednoznacznie stwierdzić, że prognoza oparta o sztuczne sieci neuronowe okazała się dokładniejsza, gdyż wskaźnik $R^2$ wyniósł dla liczby i gęstości zabitych odpowiednio: 0,64 i 0,47.

Podsumowując, ocena uzyskanych wyników w bardzo dużym stopniu zależy od przyjętych wskaźników. Biorąc jednak pod uwagę wszystkie wskaźniki, można powiedzieć, że w wielu przypadkach osiągnięto wyniki zadowalające. Przy takim podejściu łatwiejsze okazują się prognozy liczby i gęstości wypadków / zabitych, niż prognozy ich koncentracji. Niezwykle istotny jest też fakt, iż osiągnięto prognozy dokładniejsze, niż w przypadku analogicznego modelu opartego na metodzie regresji. W związku z tym można stwierdzić, iż sztuczne sieci neuronowe są narzędziem, które może być pomocne przy prognozie wskaźników bezpieczeństwa na odcinkach dróg jednojezdniowych.

### *6.2. Określenie wpływu różnych cech odcinka drogi na bezpieczeństwo na podstawie wybranych wariantów modelu.*

Po stworzeniu i zbadaniu działania wielu modeli opartych o sztucznego sieci neuronowe, wybrano optymalne dla przewidywania poszczególnych zmiennych zależnych. Poddając je dalszej analizie, można było spełnić kolejny ważny cel pracy, jakim było określenie wpływu różnych cech drogi na bezpieczeństwo ruchu.

Sposób działania w tej kwestii podał Zheng [8], a opisano go w punkcie 3.3.5. Polega on na zmianie wartości cech niezależnych o określoną wielkość i porównywaniu wynikłych z tego zmian funkcji celu (w tym przypadku zmiennej prognozowanej). Im większa będzie ta zmiana, tym istotniejsza jest dana cecha.

Do analizy wybrano sześć modeli dotyczących wypadków, określonych jako najlepsze. Są to więc następujące warianty: 23, 103, 104, 257, 275, 283. Opis ich struktury można odczytać z tabeli 5.6. W sposób optymalny prognozują one kolejno liczbę wypadków, koncentrację wypadków, koncentrację zabitych, liczbę zabitych, gęstość wypadków i gęstość zabitych, a więc wszystkie podstawowe wskaźniki bezpieczeństwa. Część spośród wybranych modeli opiera się na wszystkich cechach

wejściowych, natomiast prognozy liczby zabitych i gęstości odbywają się na podstawie wybranych 14 wejść, zgodnie z nazewnictwem zawartym w podpunkcie 5.1.1.: NAT, UC, PP, P2, OZ, PUS, PUW, PBG, WZ, SK, Z (wskaźnik gęstości zjazdów), CPR, DR, GF.

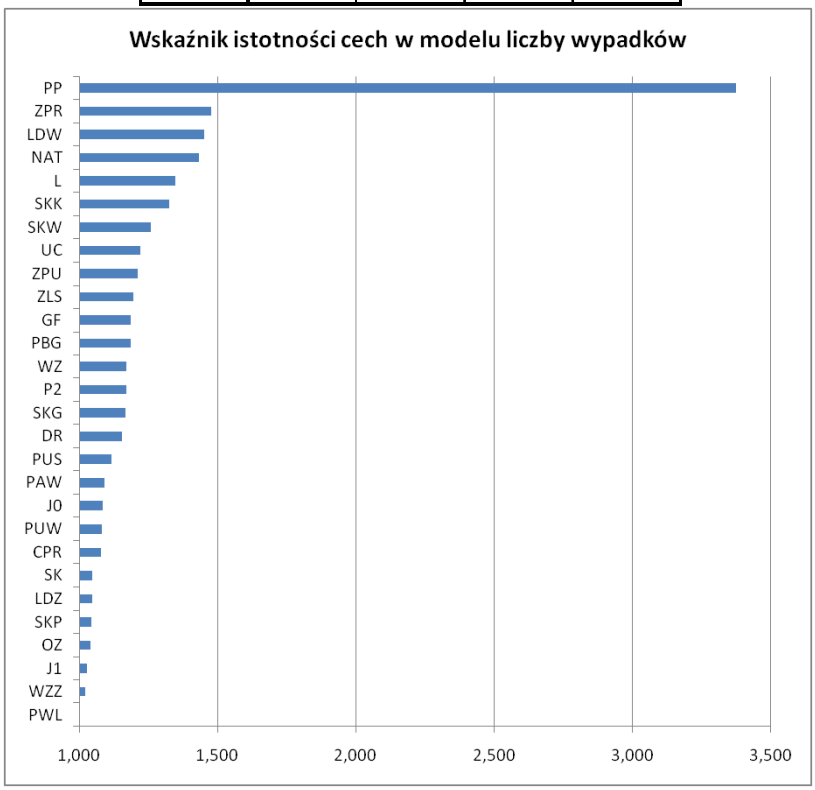
Badanie wpływu cech przeprowadzono zmieniając wartości jednej z cech objaśniających o: -σ, +σ, +2σ, +3σ, gdzie σ oznacza wartość odchylenia standardowego cechy. Zmian tych dokonywano wyłącznie na zbiorze testowym. W badaniu korzystano z nauczonej sieci neuronowej o określonej już strukturze i wagach przypisanym poszczególnym połączeniom. Dzięki temu wyeliminowano wpływ losowości na wyniki. Przy kolejnych próbach, zapisywano względną wartość zmian przewidywanej zmiennej zależnej, w stosunku do wartości uzyskanych przy oryginalnych wartościach zmiennych niezależnych. Wyniki dla poszczególnych cech i modeli przedstawiono w tabelach 6.2. – 6.7. Ponadto, utworzono wskaźnik istotności poszczególnych cech. Powstał on przez dodanie modułów zmian funkcji celu przy każdej z czterech modyfikacji zmiennej objaśniającej. Obrazuje on więc bezwzględny wpływ cechy, bez rozróżniania korelacji dodatniej i ujemnej. Wielkość wskaźnika dla poszczególnych cech w wybranych modelach przedstawiono na rysunkach 5.1. – 5.6.

Dodatkowo, analizowano również sumę wag przypisanych do określonych cech wejściowych w poszczególnych modelach. Są one zapisane dla każdego z optymalnych modeli i dostępne poprzez interfejs Scilab. Każdorazowo należało pominąć pierwszą z wag dla każdego neuronu. Stanowi ona wartość biasu, czyli dodatkowego wejścia wpływającego na sposób aktywacji neuronu [24].

Pierwszy z optymalnych modeli, prognozujący liczbę wypadków obejmuje wszystkie cechy, stąd daje ich pełne porównanie. Jak widać w poniższej tabeli, zdecydowanie najistotniejsza okazuje się być praca przewozowa (PP). Oznacza to, iż właśnie na podstawie wielkości tej cechy, w dużym stopniu opiera się prognoza liczby wypadków. Korelacja z pracą przewozową jest dodatnia, co oznacza, że jej wzrost powoduje również wzrost liczby wypadków. Należy zauważyć, że PP niesie informacje zarówno o długości odcinka, jak i natężeniu ruchu. W związku z tym, w przypadku bezwzględnego wskaźnika, jakim jest liczba wypadków, znaczny wpływ pracy przewozowej jest zgodny z oczekiwaniami [6, 22].

Tabela 6.2. Wpływ zmiany poszczególnych cech wejściowych na zmianę zmiennej zależnej w modelu prognozującym liczbę wypadków

| Cecha | Zmiana f.c. w zależności od zmiany cechy | | | |
|---|---|---|---|---|
| | -σ | +σ | +2σ | +3σ |
| PP | -9,92% | 47,66% | 78,54% | 101,34% |
| LDW | -7,26% | 7,84% | 13,12% | 16,90% |
| L | -3,64% | 6,11% | 10,71% | 14,28% |
| PBG | -3,42% | 3,34% | 5,23% | 6,46% |
| NAT | -2,20% | 9,06% | 14,24% | 17,88% |
| ZPR | -1,35% | 9,03% | 16,59% | 20,74% |
| J0 | -0,80% | 1,59% | 2,67% | 3,45% |
| DR | -0,78% | 3,44% | 5,10% | 6,04% |
| LDZ | -0,43% | 0,99% | 1,46% | 1,70% |
| SKW | -0,40% | 4,19% | 8,42% | 12,68% |
| ZPU | -0,36% | 3,61% | 7,22% | 9,99% |
| SKK | -0,31% | 5,52% | 11,03% | 15,71% |
| GF | -0,30% | 3,53% | 6,61% | 8,16% |
| OZ | -0,16% | 1,00% | 1,41% | 1,58% |
| PUW | -0,09% | 1,73% | 2,85% | 3,61% |
| SKP | -0,07% | 0,84% | 1,52% | 2,04% |
| WZZ | 0,00% | 0,36% | 0,67% | 0,95% |
| SKG | 0,00% | -3,00% | -5,69% | -8,11% |
| ZLS | 0,00% | -3,68% | -6,69% | -9,19% |
| PWL | 0,00% | 0,10% | 0,10% | 0,03% |
| PAW | 0,00% | -1,49% | -2,99% | -4,50% |
| J1 | 0,00% | 0,57% | 0,96% | 1,15% |
| WZ | 0,02% | -2,86% | -5,68% | -8,44% |
| SK | 0,07% | -0,78% | -1,54% | -2,29% |
| P2 | 0,12% | -2,73% | -5,55% | -8,45% |
| CPR | 0,35% | -1,56% | -2,58% | -3,37% |
| PUS | 0,52% | -2,51% | -3,89% | -4,77% |
| UC | 0,59% | -4,21% | -7,39% | -9,88% |




Rysunek 6.1. Wykres obrazujący wpływ poszczególnych cech wejściowych na wyniki prognoz optymalnego modelu liczby wypadków

Kolejnymi istotnymi cechami (choć już znacznie mniej od PP) wg modelu okazują się być: gęstość zjazdów prywatnych (ZPR), czy natężenie ruchu (NAT). Również wskaźnik województwa (LDW) jest znacząco powiązany z liczbą wypadków. Znikomy wpływ na funkcję celu mają z kolei m.in. udział odcinków z pasem dla ruchu powolnego (PWL), gęstość zjazdów i wjazdów na węzłach (WZZ), czy długość odcinków z drugim pasem jezdni (J1). Największą ujemną korelację można zaobserwować w przypadku udziału pojazdów ciężarowych w strukturze ruchu (UC). Wzrost tego udziału powoduje spadek liczby wypadków, co może być dość zaskakującym wnioskiem.

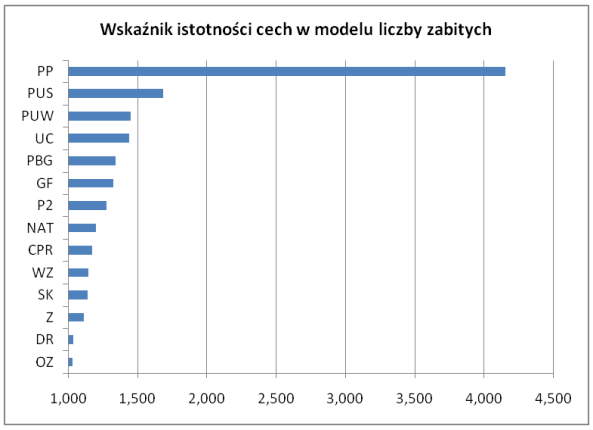
Podobne wnioski można wyciągnąć w tym przypadku z analizy wag przy odpowiednich neuronach. Zdecydowanie największa suma wag cechuje połączenia z neuronem pracy przewozowej. W dalszej kolejności należy jednak wymienić natężenie ruchu i udział odcinków z szerokim poboczem (PUW).

Przy prognozowaniu liczby zabitych, najdokładniejszy okazał się model z ograniczoną liczbą cech wejściowych. Z tego powodu nie możliwe jest pełne porównanie, jednak w zamian otrzymane dane są bardziej dokładne i lepiej oddające rzeczywisty wpływ poszczególnych cech odcinka na bezpieczeństwo.

Podobnie, jak w przypadku liczby wypadków, dla prognoz liczby zabitych decydujące znaczenie okazuje się mieć praca przewozowa. Jej wzrost wiąże się z drastycznym zwiększeniem prognoz odnośnie liczby zabitych ofiar na danym odcinku. Zaraz po PP, do istotnych cech należą PUS oraz PUW (udział odcinków z szerokim poboczem). Nie wpływają one bardzo znacząco na zwiększenie liczby wypadków, ale na liczbę zabitych już tak. Co zaskakujące, zwiększenie natężenia ruchu powoduje zmniejszenie prognozowanej liczby zabitych. Nieco wzrósł wpływ udziału samochodów ciężarowych. Niewielki wpływ mają z kolei udział odcinków zabudowanych i zadrzewionych (OZ, DR). Wielkości wag w sposób zbliżony wskazują na cechy najważniejsze i najmniej istotne.

Tabela 6.3. Wpływ zmiany poszczególnych cech wejściowych na zmianę zmiennej zależnej w modelu prognozującym liczbę zabitych

| Cecha | Zmiana f.c. w zależności od zmiany cechy | | | |
|---|---|---|---|---|
| | -σ | +σ | +2σ | +3σ |
| PP | -12,58% | 63,74% | 105,04% | 134,10% |
| PBG | -7,37% | 5,90% | 9,37% | 11,67% |
| PUS | -2,76% | 13,99% | 22,88% | 29,02% |
| UC | -1,12% | 8,21% | 14,73% | 19,98% |
| Z | -0,62% | 2,17% | 3,58% | 4,62% |
| PUW | -0,49% | 8,60% | 15,36% | 20,78% |
| DR | -0,40% | 0,81% | 1,20% | 1,42% |
| SK | -0,26% | 2,39% | 4,65% | 6,78% |
| OZ | -0,08% | 0,79% | 1,07% | 1,12% |
| WZ | 0,02% | -2,41% | -4,79% | -7,14% |
| P2 | 0,16% | -4,50% | -9,07% | -13,67% |
| GF | 0,51% | -5,94% | -11,70% | -14,62% |
| CPR | 0,82% | -3,48% | -5,67% | -7,36% |
| NAT | 1,04% | -4,14% | -6,46% | -8,08% |

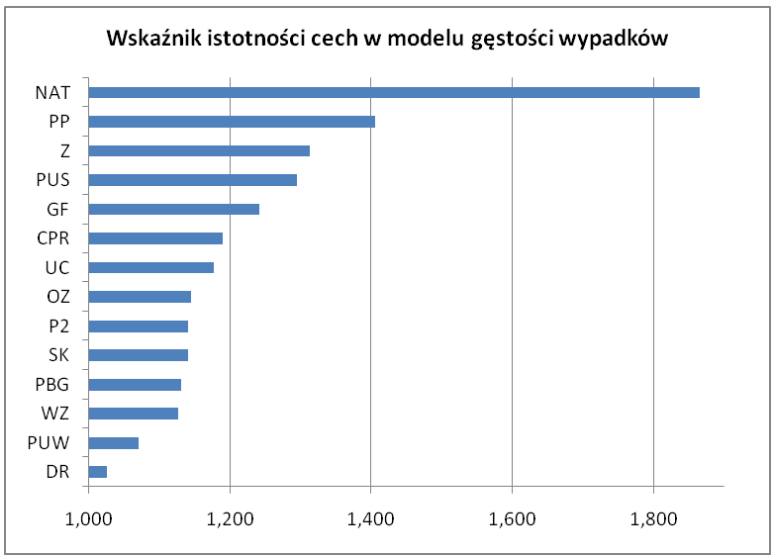



Rysunek 6.2. Wykres obrazujący wpływ poszczególnych cech wejściowych na wyniki prognoz optymalnego modelu liczby zabitych

W przypadku gęstości wypadków, korzystano z jednego najdokładniejszych uzyskanych modeli, jednak opierającego się na ograniczonej liczbie cech niezależnych. Analogiczna do powyższej analiza może w związku z tym przynieść informacje jedynie o 14 cech, wybranych na podstawie analizy dotychczasowych doświadczeń, popartej przeprowadzoną analizą PCA.

Tabela 6.4. Wpływ zmiany poszczególnych cech wejściowych na zmianę zmiennej zależnej w modelu prognozującym gęstość wypadków

| Cecha | Zmiana f.c. w zależności od zmiany cechy | | | |
|---|---|---|---|---|
| | -σ | +σ | +2σ | +3σ |
| NAT | -4,22% | 17,80% | 28,43% | 36,12% |
| PP | -2,24% | 8,74% | 13,29% | 16,34% |
| Z | -1,81% | 5,92% | 10,13% | 13,46% |
| PUS | -1,60% | 5,99% | 9,67% | 12,18% |
| CPR | -0,72% | 3,76% | 6,25% | 8,19% |
| OZ | -0,68% | 3,10% | 4,73% | 5,93% |
| GF | -0,44% | 4,46% | 8,60% | 10,68% |
| SK | -0,23% | 2,27% | 4,58% | 6,91% |
| WZ | 0,01% | -2,11% | -4,20% | -6,27% |
| P2 | 0,07% | -2,34% | -4,67% | -7,00% |
| PUW | 0,08% | -1,36% | -2,38% | -3,17% |
| DR | 0,32% | -0,45% | -0,76% | -0,99% |
| UC | 0,47% | -3,34% | -5,91% | -7,95% |
| PBG | 2,92% | -2,24% | -3,54% | -4,40% |




Rysunek 6.3. Wykres obrazujący wpływ poszczególnych cech wejściowych na wyniki prognoz optymalnego modelu gęstości wypadków

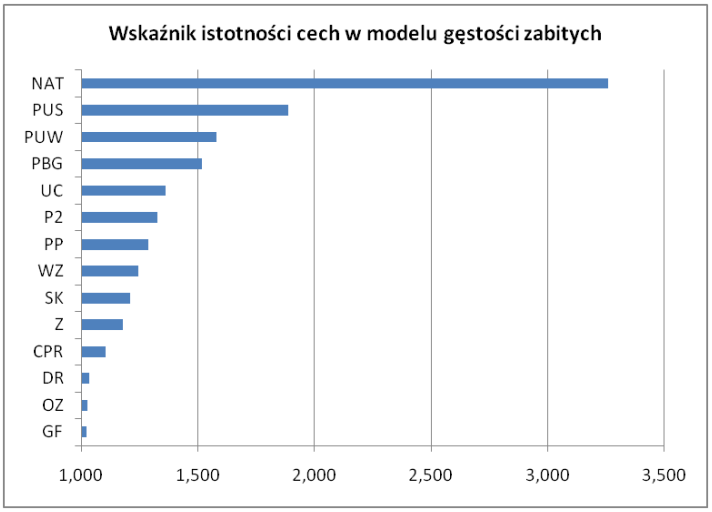
Inaczej niż w przypadku liczby wypadków, wskaźnik gęstości niweluje wpływ długości odcinka na wynika. W związku z tym, cechą najistotniejszą nie okazała się praca przewozowa, a natężenie ruchu (NAT), od którego w największym stopniu zależały wyniki prognozy. Z drugiej strony, praca przewozowa jest drugą pod względem istotności cechą, co widać na poniższym wykresie. Na trzecim miejscu znajduje się wprowadzony przez autora wskaźnik obrazujący łączną gęstość zjazdów (Z). Najmniej istotne okazują się być: udział odcinków zadrzewionych (DR) i udział odcinków z wąskim poboczem bitumicznym (PUW).

Ogólne wnioski z badania tego modelu wydają się być zbliżone do tych z modelu liczby wypadków. W modelu gęstości nie uwzględniono jednak kilku wskaźników, które poprzednio okazały się dość istotne (np. LDW, SKK, SKW), co jednak nie wpłynęło na pogorszenie jakości prognoz. Poza tym, zwraca uwagę fakt, że w przypadku wskaźnika dotyczącego udziału odcinków z poboczem gruntowych (PBG), w jednym przypadku zanotowano korelację dodatnią, a w drugim ujemną.

Jeśli chodzi o wagi w sieci neuronowej, największe występują przy natężeniu ruchu, jednak najmniejsze przy pracy przewozowej. Tak znaczna różnica w stosunku do opisanej wyżej metody może wskazywać, że bezpośrednia analiza wag nie jest w pełni wiarygodna.

Tabela 6.5. Wpływ zmiany poszczególnych cech wejściowych na zmianę zmiennej zależnej w modelu prognozującym gęstość zabitych

| Cecha | Zmiana f.c. w zależności od zmiany cechy | | | |
|---|---|---|---|---|
| | -σ | +σ | +2σ | +3σ |
| PBG | -10,81% | 9,07% | 14,27% | 17,66% |
| NAT | -9,50% | 45,10% | 75,03% | 96,22% |
| PUS | -3,67% | 17,76% | 29,52% | 37,79% |
| PP | -1,54% | 6,15% | 9,39% | 11,56% |
| Z | -1,11% | 3,43% | 5,82% | 7,71% |
| UC | -0,89% | 6,61% | 12,03% | 16,52% |
| PUW | -0,63% | 10,97% | 19,67% | 26,66% |
| DR | -0,41% | 0,66% | 1,07% | 1,34% |
| SK | -0,36% | 3,38% | 6,85% | 10,40% |
| WZ | -0,02% | 3,97% | 8,07% | 12,27% |
| GF | 0,04% | -0,44% | -0,85% | -1,06% |
| OZ | 0,14% | -0,57% | -0,86% | -1,08% |
| P2 | 0,17% | -5,59% | -10,93% | -16,02% |
| CPR | 0,45% | -2,09% | -3,41% | -4,43% |




Rysunek 6.4. Wykres obrazujący wpływ poszczególnych cech wejściowych na wyniki prognoz optymalnego modelu gęstości zabitych

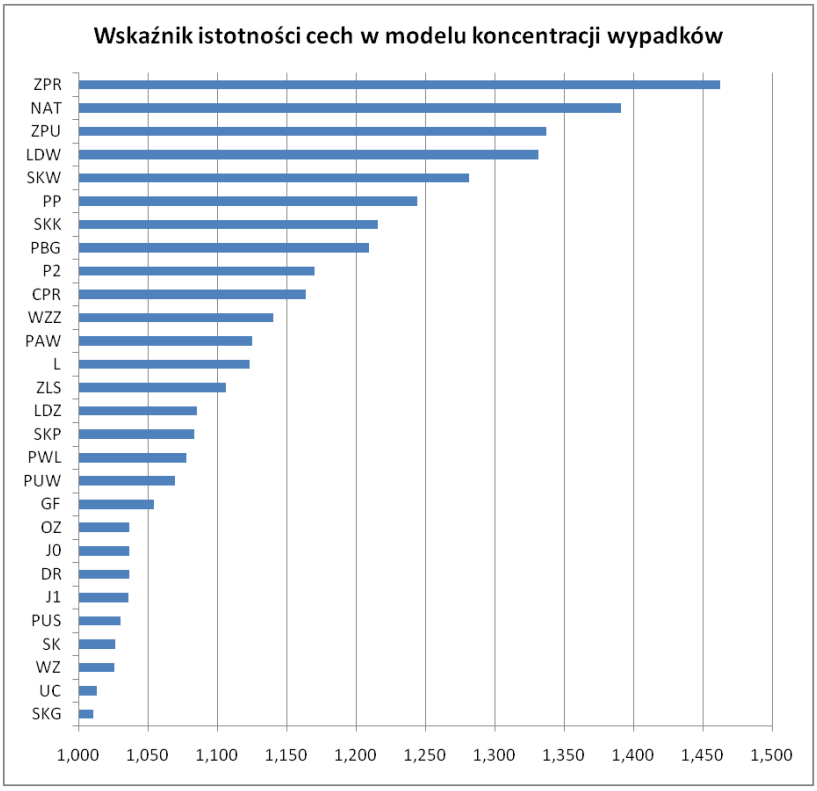
W modelu gęstości zabitych również kluczowe okazuje się być natężenie ruchu. W dalszej kolejności model wykazuje podobieństwo do modelu liczby zabitych, ze względu na istotność PUS, PUW, czy UC. Podobieństwo jest widoczne również w przypadku najmniej znaczących cech (OZ, DR). Największa ujemna korelacja cechuje wskaźnik PBG, co oznacza, że zwiększenie długości odcinków z poboczem gruntowym wpływa pozytywnie na zmniejszenie liczby ofiar zabitych.

W opisanym powyżej modelu gęstości zabitych, wagi przy poszczególnych neuronach w sposób niemal idealny odwzorowują wyniki analizy wpływu cech wejściowych na wyniki prognoz.

Kolejnym z przeanalizowanych modelów był ten dotyczący koncentracji wypadków. Zbudowany był on w oparciu o wszystkie dostępne cechy.

Tabela 6.6. Wpływ zmiany poszczególnych cech wejściowych na zmianę zmiennej zależnej w modelu prognozującym koncentrację wypadków

| Cecha | Zmiana f.c. w zależności od zmiany cechy | | | |
|---|---|---|---|---|
| | -σ | +σ | +2σ | +3σ |
| LDW | -5,02% | 5,96% | 9,78% | 12,41% |
| PBG | -4,04% | 3,62% | 5,88% | 7,43% |
| L | -1,59% | 2,13% | 3,71% | 4,93% |
| ZPR | -1,33% | 8,66% | 16,07% | 20,16% |
| ZPU | -0,52% | 5,46% | 11,04% | 16,72% |
| DR | -0,46% | 0,65% | 1,10% | 1,43% |
| SKW | -0,45% | 4,53% | 9,19% | 13,97% |
| PUS | -0,33% | 0,59% | 0,94% | 1,18% |
| SKK | -0,20% | 3,55% | 7,25% | 10,54% |
| OZ | -0,15% | 0,75% | 1,24% | 1,54% |
| PUW | -0,04% | 1,40% | 2,40% | 3,14% |
| SK | -0,03% | 0,41% | 0,87% | 1,36% |
| PWL | 0,00% | 1,41% | 2,63% | 3,70% |
| SKG | 0,00% | 0,18% | 0,36% | 0,54% |
| ZLS | 0,00% | -2,00% | -3,63% | -4,97% |
| WZ | 0,00% | -0,43% | -0,86% | -1,28% |
| PAW | 0,00% | -2,10% | -4,18% | -6,25% |
| J1 | 0,02% | -0,59% | -1,18% | -1,77% |
| WZZ | 0,04% | -2,40% | -4,70% | -6,92% |
| UC | 0,04% | -0,27% | -0,44% | -0,55% |
| P2 | 0,08% | -2,92% | -5,70% | -8,33% |
| GF | 0,12% | -1,11% | -1,92% | -2,28% |
| SKP | 0,14% | -1,38% | -2,74% | -4,06% |
| J0 | 0,53% | -0,63% | -1,08% | -1,42% |
| CPR | 0,68% | -3,35% | -5,40% | -6,94% |
| LDZ | 1,12% | -1,63% | -2,58% | -3,20% |
| PP | 1,59% | -5,44% | -7,93% | -9,48% |
| NAT | 2,19% | -7,85% | -13,24% | -15,79% |



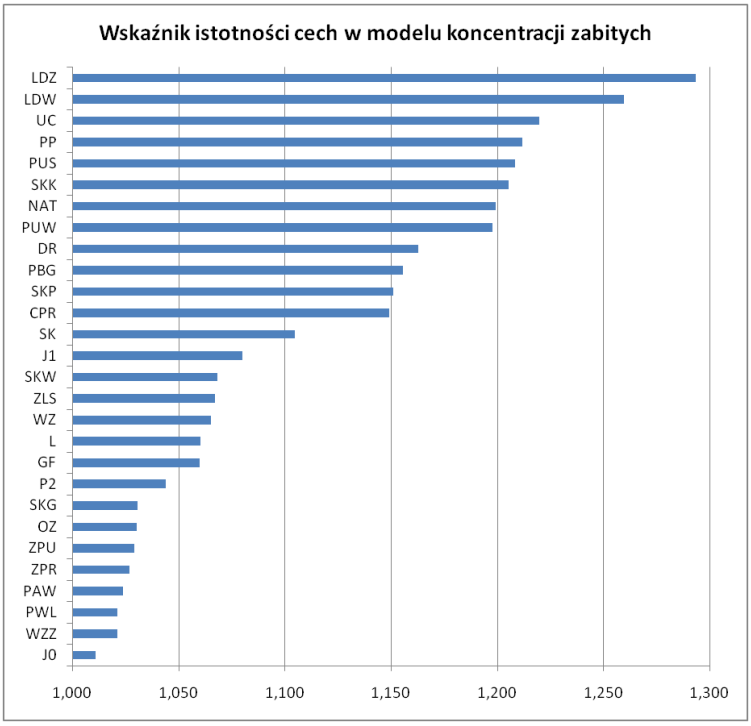

Rysunek 6.5. Wykres obrazujący wpływ poszczególnych cech wejściowych na wyniki prognoz optymalnego modelu koncentracji wypadków

Wskaźnik koncentracji pozwala wyeliminować wpływ zarówno długości badanego odcinka, jak i natężenia ruchu na nim. W efekcie, w stosunku do poprzednich modeli, znacznie zmalało znaczenie wskaźnika pracy przewozowej, a także natężenia ruchu i długości odcinka. Wpływ poszczególnych cech w tym modelu jest znacznie bardziej zrównoważony. Najistotniejsza okazuje się być gęstość zjazdów prywatnych (ZPR). Jej wzrost w znaczący sposób sprzyja powstawaniu wypadków, podobnie jest w przypadku zjazdów publicznych.

Bardzo odmiennie kształtuje się wpływ czynników na wskaźnik koncentracji ofiar zabitych. Najistotniejsze okazują się być LDZ i LDW (wskaźniki województw) oraz L, czyli długość odcinka, które dodatkowo cechuje silna korelacja ujemna. ZPR, który w przypadku koncentracji wypadków należał do najistotniejszych czynników, przy ofiarach zabitych okazuje się mieć niewielkie znaczenie, podobnie jak ZPU. Spadło również nieco znaczenie wielkości natężenia ruchu (NAT), natomiast istotność pracy przewozowej utrzymała się na podobnym poziomie. Co warte odnotowania, wpływ udziału samochodów ciężarowych (UC) drastycznie wzrósł, stając się jednym z najistotniejszych czynników. Wskaźnik zalesienia (ZLS) pozostał wśród czynników mało znaczących dla zmiany koncentracji wypadków na odcinku drogi.

Tabela 6.7. Wpływ zmiany poszczególnych cech wejściowych na zmianę zmiennej zależnej w modelu prognozującym koncentrację zabitych

| Cecha | Zmiana cechy (liczba odchyleń standardowych) | | | |
|---|---|---|---|---|
| | -σ | +σ | +2σ | +3σ |
| LDW | -4,31% | 4,63% | 7,53% | 9,50% |
| L | -3,93% | -1,09% | 0,07% | 0,93% |
| LDZ | -3,61% | 5,39% | 8,93% | 11,42% |
| PBG | -2,65% | 2,86% | 4,50% | 5,57% |
| DR | -1,76% | 3,12% | 5,04% | 6,34% |
| PUS | -1,53% | 4,26% | 6,72% | 8,33% |
| UC | -0,57% | 4,12% | 7,35% | 9,93% |
| SKP | -0,23% | 2,52% | 4,98% | 7,39% |
| PUW | -0,22% | 3,86% | 6,74% | 8,95% |
| J0 | -0,16% | 0,20% | 0,32% | 0,41% |
| OZ | -0,10% | 0,70% | 1,04% | 1,21% |
| ZPR | -0,07% | 0,57% | 0,94% | 1,12% |
| WZZ | -0,01% | 0,37% | 0,71% | 1,02% |
| WZ | -0,01% | 1,09% | 2,18% | 3,26% |
| SKG | 0,00% | -0,54% | -1,04% | -1,50% |
| ZLS | 0,00% | 1,50% | 1,50% | 3,73% |
| PWL | 0,00% | 0,43% | 0,74% | 0,96% |
| PAW | 0,00% | -0,38% | -0,79% | -1,23% |
| P2 | 0,02% | -0,69% | -1,44% | -2,26% |
| J1 | 0,04% | -1,31% | -2,64% | -4,01% |
| ZPU | 0,05% | -0,44% | -0,93% | -1,48% |
| SKW | 0,12% | -1,09% | -2,22% | -3,41% |
| GF | 0,13% | -1,00% | -2,13% | -2,74% |
| SK | 0,19% | -1,66% | -3,40% | -5,22% |
| SKK | 0,22% | -3,44% | -6,94% | -9,94% |
| CPR | 0,64% | -2,99% | -4,90% | -6,37% |
| NAT | 1,09% | -3,91% | -6,77% | -8,16% |
| PP | 1,36% | -4,67% | -6,87% | -8,29% |




Rysunek 6.6. Wykres obrazujący wpływ poszczególnych cech wejściowych na wyniki prognoz optymalnego modelu koncentracji zabitych

W obydwu modelach koncentracji, wagi przy połączeniach dość znacznie różnią się od zaobserwowanego rzeczywistego wpływu cech na wyniki.

Podsumowując, wpływ poszczególnych cech na badaną wielkość (np. liczbę, gęstość wypadków), określano trzystopniowo:

- po pierwsze, wyciągnięto wnioski z analizy dotychczasowych doświadczeń w dziedzinie wykorzystania sztucznych sieci neuronowych do badań wypadków drogowych. Na tej podstawie dobrano zestawy cech, stanowiących warstwę wejściową kolejnych modeli,
- po drugie, przeprowadzono analizę PCA, która potwierdziła słuszność wcześniejszych wniosków, jednocześnie dostarczając dodatkowego materiału do analizy,
- po trzecie, po zbudowaniu modeli i przeprowadzeniu prognoz, dokonano ich analizy pod kątem istotności wpływu cech na zachowaniu modelu.

Wszystkie trzy etapy wskazują, że decydujący wpływ na liczbę i gęstość wypadków mają praca przewozowa i natężenie ruchu. Wpływ ten dość mocno eliminuje zastosowanie wskaźnika koncentracji wypadków. Dotychczasowe badania potwierdzały duże znaczenie udziału pojazdów ciężarowych w strukturze ruchu [6].

Tymczasem zarówno analiza PCA, jak i wyniki badań modeli wykorzystujących sieci neuronowe, wskazują raczej na jego niewielką, a w przypadku koncentracji wypadków wręcz znikomą istotność w przypadku wypadków. W modelach dotyczących ofiar zabitych, UC jest już ważniejszym wskaźnikiem, którego wzrost powoduje pogorszenie bezpieczeństwa na danym odcinku. Z kolei dla istniejącej bazy danych dotyczących sieci dróg krajowych, silnie skorelowane z wypadkami okazują się być wskaźniki dotyczące liczby zjazdów. Według analizy PCA i modelu z sieciami neuronowymi, ich duża gęstość powoduje wzrost liczby, gęstości i koncentracji wypadków.

Potwierdza się prawidłowość dotycząca wpływu lokalizacji odcinka drogi na bezpieczeństwo ruchu. Wniosek taki można wysnuć na podstawie wyników analiz dla wskaźnika bezpieczeństwa województw (LDW), którego wzrost dość mocno wiąże się ze wzrostem wskaźników wypadków. Inaczej jest w przypadku długości odcinków z poboczem gruntowym. Wskaźnik ten okazuje się być dość istotny i w przypadku analizy PCA, modelu gęstości i koncentracji jego wzrost powoduje zmniejszenie wskaźników dotyczących wypadków. Jedynie w modelu liczby wypadków, korelacja tego wskaźnika jest dodatnia.

Wszystkie modele potwierdzają natomiast znikomy wpływ wskaźników dotyczących liczby zjazdów na węzłach, a zwłaszcza wskaźnika zadrzewienia. Oznacza to, iż obecność drzew przy drodze praktycznie nie ma wpływu na wskaźniki dotyczące wypadków. Podobnie jest w przypadku liczby, gęstości, koncentracji zabitych – zadrzewienie nie ma wielkiego znaczenia. Wskaźniki te cechuje z kolei duża istotność udziału odcinków z poboczem bitumicznym (zarówno szerokim, jak i wąskim). W przeciwieństwie do pobocza gruntowego, jego występowanie sprzyja wzrostowi wymienionych wskaźników.

Na koniec można również stwierdzić, że ze względu na złożoność struktury sieci neuronowej, nie zawsze wartości wag połączeń przekładają się bezpośrednio na wyniki analiz osiągnięte przy innym sposobie badania wpływu czynników. W większości przypadków suma wag była największa dla rzeczywiście najważniejszej cechy, jednak już dalsza kolejność różniła się znacząco. Największą zbieżność wyników i wag zaobserwowano w przypadku sieci prognozującej gęstość zabitych, która posiadała najprostszą strukturę.

# 7. PODSUMOWANIE.

Celem pracy było stworzenie modelu, wykorzystującego sztuczne sieci neuronowe i prognozującego wskaźniki bezpieczeństwa na odcinkach dróg krajowych w Polsce. Poza tym, przy użyciu modelu należało określić wpływ poszczególnych czynników na powstawanie wypadków. Cele te udało się osiągnąć.

Na początku zapoznano się z literaturą dotyczącą w ogólny sposób sztucznych sieci neuronowych, a także zagadnienia modelowania bezpieczeństwa na polskich drogach krajowych. W sposób szczególny odwołano się również do polskiego modelu prognoz bezpieczeństwa na odcinkach dróg krajowych z wykorzystaniem metody regresji. Stanowił on punkt wyjścia do tworzenia modelu wykorzystującego sieci neuronowe.

Następnie omówiono szczegółowo zagraniczne modele, krok po kroku przedstawiając procedurę badawczą. Spośród licznych dotychczasowych prac Skupiono się na tych, które mogły być użyteczne przy tworzeniu określonego w założeniach modelu, tj. w sposób szczególny dotyczyły prognoz liczby (lub gęstości) zdarzeń na odcinkach dróg. Przedstawiono sposoby przygotowywania danych, budowania modeli, doboru zmiennych wejściowych i wyjściowych, testowania i oceny wyników. Omówiono również stosowaną przez badaczy metodę Analizy Głównych Składowych (PCA), która jest pomocna przy określaniu istotności i wpływu poszczególnych cech odcinka na bezpieczeństwo.

Po zgromadzeniu wiedzy teoretycznej wiedzy dotyczącej badanego zagadnienia, przystąpiono do budowy własnego modelu. Zgodnie z uściślonymi w punkcie 4. założeniami, skorzystano z gotowej bazy danych odcinków dróg krajowych jednojezdniowych. Jako element prognozowany przyjęto takie wskaźniki, jak: liczba, gęstość i koncentracja wypadków oraz zabitych. Liczbę i rodzaj zmiennych wejściowych określono wzorując się na opisanych wcześniej badaniach, przeprowadzono również w tym celu analizę PCA. Ostatecznie do testowania wybrano pięć różnych zestawów cech opisujących każdy odcinek.

Jako narzędzie modelowania wybrano środowisko Scilab wraz ze specjalnym Toolboxem przeznaczonym do tworzenia sieci neuronowych [23]. Wybór ten pozwolił na łatwe zmienianie poszczególnych cech modelu i testowaniu licznych jego wariantów. Ostatecznie przetestowano ponad 400 różnych wariantów różniących się zestawami zmiennych wejściowych, wyjściowych, liczbą epok uczenia i strukturą wewnętrzną sieci. Proces testowania przedstawiono na przykładzie modelu z 28 zmiennymi wejściowymi.

Na podstawie przyjętych kryteriów oceny (analogicznych do stosowanych w dotychczasowych badaniach) wybrano te, które najlepiej prognozowały najważniejsze wskaźniki bezpieczeństwa. Ich strukturę oraz osiągane wskaźniki dokładności prognoz przedstawiono w tabelach. Dodatkowo, podjęto próbę poprawy dokładności jednego z wariantów modelu poprzez zastosowanie zmienionej metody doboru liczby epok uczenia, a także zmieniania liczby danych uczących.

Otrzymane wyniki przeanalizowano i porównano z innymi badaniami. Na podstawie tej analizy oraz w odniesieniu do przedstawionej literatury, można wyciągnąć następujące, najważniejsze wnioski:

- sztuczne sieci neuronowe są narzędziem, które można wykorzystać do prognozowania wskaźników bezpieczeństwa na drogach,

- dokładność prognoz zależy od rodzaju prognozowanych wskaźników bezpieczeństwa oraz od struktury sieci, użytych danych wejściowych i innych czynników,

- ocena dokładności może być różna w zależności od przyjętych wskaźników oceny,

- możliwe jest uzyskanie prognoz, których dokładność jest zadowalająca i konkurencyjna w stosunku do innych metod prognozowania, takich jak regresja wieloraka. Ocenę dokładności prognoz stworzonego modelu przedstawiono szerzej w punkcie 5.1.

- uzyskano zasadniczo zbieżne z dotychczasowymi doświadczeniami efekty, jeżeli chodzi o określenie najistotniejszych czynników wpływających na liczbę, gęstość i koncentrację wypadków. Są to m.in. natężenie ruchu, praca przewozowa oraz gęstość zjazdów. Szerokie omówienia tej kwestii umieszczono w punkcie 5.2.

- analiza PCA jest badaniem, które może być pomocne przy tworzeniu modelu opartego o sztuczne sieci neuronowe. Z drugiej strony może stanowić odrębne narzędzie badań i być wykorzystane niezależnie od sieci neuronowych, przynosząc wiedzę odnośnie istotności i wpływu cech odcinków na bezpieczeństwo ruchu drogowego.

Jednym z najważniejszych rezultatów pracy są modele prognozujące określone wskaźniki wypadków. Przygotowano je w taki sposób, aby mogły one zostać wykorzystane do dalszych badań. W tym celu, zachowano strukturę pięciu optymalnych wariantów, służących prognozowaniu:

- liczby wypadków (dodatkowo liczby rannych, ciężko rannych i zabitych),

- gęstości wypadków,

- koncentracji wypadków (dodatkowo koncentracji rannych, ciężko rannych i zabitych).

W modelach tych, oprócz struktury sieci neuronowej, zapamiętano również wagi połączeń między warstwami uzyskane w procesie uczenia sieci. Dzięki temu w szybki sposób, można przeprowadzić prognozę dla odcinków dróg o dowolnych cechach. Ważne jest, aby dostarczyć odpowiednią ilość danych o strukturze dokładnie takiej, jaką posiadały dane wykorzystane pierwotnie. W ten sposób można dokonywać prognoz m.in. dla istniejących odcinków o zmienionym natężeniu ruchu lub innych parametrach, nowopowstałych odcinków dróg, odcinków dróg określonych w wyniku innego niż zaproponowany podziału. Modele wraz z opisem sposobu korzystania z nich umieszczono w załącznikach.

Oprócz zbudowania modeli, efektem procesu badawczego jest szereg obserwacji i wytycznych, które mogę być pomocne przy tworzeniu zupełnie odrębnych modeli dotyczących bezpieczeństwa ruchu na odcinkach dróg. Poniżej przedstawiono najistotniejsze kwestie, jakie należy wziąć pod uwagę tworząc taki model, uszeregowane w kolejności zgodnej z ich zaistnieniem w procesie badawczym:

- do budowy modelu bezpieczeństwa opartego na sztucznych sieciach neuronowych potrzebna jest baza danych zawierająca kilkaset rekordów,

- każdy z rekordów powinien stanowić odcinek opisany przez co najmniej kilkanaście cech, spośród których najważniejsze są praca przewozowa, natężenie ruchu, gęstość zjazdów,

- nowo tworzony model powinien być osobno testowany przy zmiennych parametrach, takich jak liczba epok uczenia, liczba danych wejściowych, struktura sieci neuronowej i innej. Na podstawie wniosków z niniejszej pracy nie udało się określić szczegółowych zasad doboru wymienionych wyżej cech do poszczególnych modeli. Ogólnie można przyjąć, że im mniej skomplikowany ma być model (większa liczba cech wyjściowych i wejściowych), tym bardziej należy uprościć jego strukturę,

- zwiększanie liczby cech niezależnych nie gwarantuje uzyskania dokładniejszych prognoz, przeciwnie, może powodować pogorszenie osiąganych rezultatów,

- testowanie modelu powinno opierać się na przyjętych wskaźnikach, takich jak MRE, MSE, $r^2$. Monitorowanie tych wskaźników dokładności prognoz modelu prowadzi do określenia optymalnej struktury.

## WYKAZ LITERATURY.


1. Kowalski K., Kaplar I., Maśkiewicz J.: *Podstawowe statystyki wypadków drogowych na zamiejskiej sieci dróg krajowych 2011*. Wydział Pomiarów Ruchu Departament Studiów GDDKiA, czerwiec 2012.

2. Nowakowska M.: *Analiza typologiczna wypadków drogowych z wykorzystaniem sztucznej sieci neuronowej Kohonena.* Drogownictwo, t. 10, 2002, s. 333 – 339.

3. Ambroch K.: *Sztuczne sieci neuronowe*. Matematyka – Społeczeństwo – Nauczanie, nr 32, s. 44 – 47, ISSN 1427 – 1591.

4. Karlaftis M.G., Vlahogianni E.I.: *Statistical methods versus neural network in transportation research: Differences, similarities and some insights*. Transportation research, nr 19, 2011, s. 387 – 399.

5. Isaak S., Trifiro F.: *Artificial Intelligence In Transportation. Neural Networks.* Transport research circular, nr E-C113, 2007, s. 17 – 32.

6. Kustra W., Jamroz K.: *Analiza czynników wpływających na gęstość ofiar śmiertelnych na drogach krajowych w Polsce*. Journal of KONBiN, nr 1, 2010, s. 221 – 234.

7. Akiyama T., Kotani Y., Suzuki T.: *The Optimal Transport Safety Planning with Accident Estimation Process.*

8. Zheng L. Meng X.: *An approach to predict road accident frequencies: application of fuzzy neural Network.* 3rd International Conference on Road Safety and Simulation, Indianapolis, USA, Wrzesień 2011.

9. Bosurgi G., D'Andrea A, Trifirò F.: *A Motorway Safety Analysis Model Based on Artificial Neural Networks.*

10. Rezaie Moghaddam F. Afandizadeh Sh., Ziyadi M.: *Prediction of accident severity using artificial neural networks.* International Journal of Civil Engineering, nr 1, 2011, s. 41 – 48.

11. Miao M. C., Abraham A., Paprzycki M.: *Traffic accident analysis using decision trees and neural networks*. Computer Science Department, Oklahoma State University, USA.

12. Shanthi S., Geetha Ramani R.: *Classification of Vehicle Collision Patterns in Road Accidents using Data Mining Algorithms.* International Journal of Computer Applications, nr 12(35), 2011, s. 30 – 37.

13. Sohn S.Y., Shin H.: *Pattern recognition for road traffic accident severity in Korea.* Ergonomics. nr 44(1), 2001, s. 107-117.

14. Hosseinpour M., Yahaya A.S., Ghadiri S.M., Prasetijo J.: *Application of Adaptive Neuro-Fuzzy Inference System for Road Accident Prediction*. International Journal of Civil Engineering, nr 17(7), 2013, s. 1761-1772.

15. Tadeusiewicz R., *Sieci neuronowe.* Warszawa, Akademicka Oficyna Wydaw. RM, 1993.

16. Chang L.-Y.: *Analysis of freeway accident frequencies: Negative binomial regression versus artificial neural network.* Safety Science, nr 43, 2005, s. 541 – 557.

17. *Mining road traffic accidents*. University of Jyvaskyla, Finlandia, Software and Computational Engineering, ISBN 9789513937522.

18. Gramacki J., Gramacki A: *Redukcja wymiarowości oraz wizualizacja danych wielowymiarowych z wykorzystaniem projektu R*. XV Konferencja PLOUG, Kościelisko, Październik 2009.

19. Koronacki J., Mielniczuk J.: *Statystyka dla studentów kierunków technicznych i przyrodniczych*. Warszawa: WNT, 2006, s. 41-46, 270-272. ISBN 83-204-3242-1.

20. Nowakowska M.: *Zaawansowane metody w badaniach i modelowaniu bezpieczeństwa ruchu drogowego.* IX Międzynarodowe Seminarium Bezpieczeństwa Ruchu Drogowego GAMBIT 2012, Gdańsk, kwiecień 2012.

21. Bishop, C.: *Neural Networks for Pattern Recognition.* Oxford University Press, 1995, s.116-149.

22. Budzyński M., Kustra W.: *Analiza zagrożeń na jednorodnych odcinkach dróg*. Drogownictwo, nr 4, 2012.

23. Hristev R.M., Cornet A.: ANN Toolbox, wersja: 0.4.2.5-2, 24.11.2011, strona internetowa: https://atoms.scilab.org/toolboxes/ANN_Toolbox; dostęp: 20.08.2014.

24. Bartkowski B., Przydróżny M.: *Sztuczne sieci neuronowe*, strona internetowa: http://sknbo.ue.poznan.pl/neuro/ssn/pliki/uczenie/uczenie1.html; dostęp: 26.08.2014.

## WYKAZ RYSUNKÓW

## WYKAZ TABEL